UNIVERSIDADE DE LISBOA

FACULDADE DE PSICOLOGIA

ESTUDO DE EQUIVALÊNCIA INTERCULTURAL DA BATERIA DE RACIOCÍNIO CRÍTICO (*CRTB*) EM CINCO PAÍSES COM RECURSO AO MODELO DE RASCH (TRI)

Tiago Alexandre Monteiro Rebelo Moreninho

MESTRADO INTEGRADO EM PSICOLOGIA

(Secção de Psicologia dos Recursos Humanos, do Trabalho e das Organizações)

2011

UNIVERSIDADE DE LISBOA

FACULDADE DE PSICOLOGIA

ESTUDO DE EQUIVALÊNCIA INTERCULTURAL DA BATERIA DE RACIOCÍNIO CRÍTICO (*CRTB*) EM CINCO PAÍSES COM RECURSO AO MODELO DE RASCH (TRI)

Tiago Alexandre Monteiro Rebelo Moreninho

MESTRADO INTEGRADO EM PSICOLOGIA

(Secção de Psicologia dos Recursos Humanos, do Trabalho e das Organizações)

Dissertação Orientada pela Professora Doutora Maria João Afonso

2011

Agradecimentos

E chegou ao fim…

Foi uma etapa muito trabalhosa, cheia de contratempos, descobertas pessoais e profissionais. Durante estes meses senti alegria e tristeza, calma e ansiedade, euforia e frustração. Agora, ao cruzar esta meta, olho para trás e vejo que, ao lado dos meus passos, estão muitos outros das pessoas que me acompanharam, a quem agora passo a agradecer:

À Professora Doutora Maria João Afonso, por ter navegado comigo nesta ambiciosa e ousada viagem. A sua experiência, conhecimento, seriedade, e o seu rigor científico e metodológico, validaram todas as escolhas que efectuei neste percurso, e foram cruciais para que o presente trabalho ficasse concluído. Não posso deixar de agradecer todo o apoio que me deu, tanto a nível pessoal como profissional, e reservo-me o direito de elogiar a sua capacidade de me levantar a moral nos períodos mais difíceis. Um sincero Muito Obrigado por me ter deixado aprender tanto consigo.

Aos Professores da Secção de Psicologia dos Recursos Humanos, do Trabalho e das Organizações, pela transferência de conhecimentos que fizeram nos últimos dois anos e pelos desafios que me colocaram. Quero também deixar uma palavra de agradecimento, pela abertura e flexibilidade demonstrada na colaboração da Professora Maria João Afonso com a Secção.

Ao Dr. Jorge Horta Alves e à Dra. Isabel Paredes, por terem proporcionado a realização desta dissertação disponibilizando todos os recursos necessários para tal, e principalmente por todo o apoio demonstrado ao longo destes meses.

À Joana Macedo, pela incansável procura da obtenção dos dados, pela persistência e pela ajuda em procurar alternativas quando as portas se iam progressivamente fechando.

À Helena Monteiro, por me ter ajudado a dar o importante primeiro passo na análise de dados, por me ensinar a utilizar o avançado programa estatístico e por ainda se ter disponibilizado a verificar dados.

À minha Mãe, por todo o apoio e preocupação demonstrada. Foi indubitavelmente um ano de mudanças, pessoais e profissionais, umas mais fáceis, outras mais difíceis. A verdade é que aqui estamos, numa nova etapa da nossa vida, cujo caminho desconhecemos e nada

conseguimos anteceder, excepto uma coisa que passo a prometer: irei compensar todas as refeições *gourmet* que tive ao longo deste ano!

Ao meu Pai, pela compreensão da minha ausência nos fins-de-semana familiares, e por manter um quarto cativo na minha pensão preferida. Foi de facto um ano atribulado mas agora, tudo se avizinha mais calmo e disponível.

Às restantes pessoas da minha família, especialmente à Margarida que, entre disputas normativas entre irmãos, também ajudou a verificar dados; à Beatriz, pelos grandes abraços nas reuniões familiares e por me fazer rir quando diz que o 1º Ano a deixa muito cansada; à minha Avó, pelas saudades que teve de aguentar durante este ano; à Teresa, pelo interesse que sempre demonstrou e pelas nossas conversas na cozinha; e ao Luís, por ter sido um facilitador nestas literais mudanças. Um especial agradecimento à minha nova família Palhoco, por me terem acolhido tão bem.

Às minhas amigas dissertivas, Mafalda, Ana Rita, Catarina e Inês, por terem partilhado comigo ansiedades, frustrações e conquistas, por se preocuparem constantemente em saber o andamento deste projecto. Juntos começámos, juntos acabámos!

Um especial agradecimento à Andreia Rosa e à Leonor Horta por me inspirarem a ir mais além, a desafiar barreiras e a ultrapassar obstáculos, que me levaram a embarcar neste projecto tão ambicioso.

Aos meus amigos, colegas organizacionais e clínicos, colegas de estágio e agora, de trabalho, pelo apoio, carinho e preocupação demonstrados durante este percurso. Também aos meus colegas da dança, em especial à Teresa e à Lena, pela constante preocupação, todas as semanas, sobre o paradeiro dos meus dados e da minha tese.

Por último, à Rita. Obrigado pelo crescimento conjunto, pelo esforço em superar os desafios e as contrariedades, por ser emocional nos momentos em que estava mais racional, por me fazer sorrir quando me sentia mais triste e por me ensinar a empatizar e a reconhecer emoções, enquanto também aprendia algumas coisas de TRI. Não foi fácil fazermos a tese ao mesmo tempo, mas se superámos isto, superamos qualquer coisa!

Índice

Lista de Tabelas

Resumo

O presente trabalho de investigação aplicada centra-se na temática da equivalência intercultural, e consiste no estudo de uma bateria de testes de aptidões, utilizada em diversas organizações nacionais e internacionais, em contextos de avaliação psicológica – *Bateria de Raciocínio Crítico (CRTB)*. Partiu do pressuposto de que as versões originais e as versões adaptadas de cada teste são equivalentes, e procedeu à comparação entre as versões do Reino Unido (original) e de Portugal, da Austrália e da África do Sul (adaptações), e de Portugal (original) e Moçambique (adaptação).

Partindo de um total 4946 respostas a três testes de aptidões - Verbal, Numérico e Diagramático - foi efectuado o estudo do Funcionamento Diferencial dos Itens (*DIF*), nas versões aplicadas nos cinco países, com recurso a metodologia baseada na Teoria da Resposta ao Item (TRI), nomeadamente, pela aplicação do Modelo de Rasch.

Detectou-se a presença de *DIF* em alguns itens, nos vários testes, o que pode comprometer a equivalência intercultural das diferentes versões. Contudo, a análise efectuada não permitiu concluir que os testes são, na sua globalidade, culturalmente enviesados, devido a limitações na qualidade dos dados amostrais disponíveis.

Este estudo alertou para a necessidade de averiguar a qualidade dos conteúdos dos itens e para a conveniência da utilização conjunta de Métodos Estatísticos e dos Juízes, nas adaptações de testes, assim como para a necessidade de interpretação cautelosa, em decisões de selecção que tenham por base os resultados de instrumentos psicométricos adaptados. Adicionalmente, reuniu evidências das potencialidades e vantagens da metodologia TRI para a investigação em Psicologia Intercultural.

Palavras-chave: Bateria de Raciocínio Crítico *(CRTB)*, Equivalência Intercultural, Funcionamento Diferencial do Item (*DIF*), Teoria da Resposta ao Item (TRI), Tradução/Adaptação de testes.

Abstract

This applied research project focuses on the problem of cross-cultural equivalence and aims at studying an ability test battery, used in psychological assessments from different national and international organizations – the *Critical Reasoning Test Battery (CRTB)*. It was based on the assumption that both the original and the adapted versions must be equivalent and it compares the version used in the United Kingdom (original) with the one used in Portugal, in Australia and in South Africa (adaptations), as well as the version used in Portugal (original) with the one used in Mozambique (adaptation).

From a total of 4946 responses to three ability tests – Verbal, Numerical and Diagrammatic – the presence of Differential Item Functioning (DIF) was studied on the versions applied in five countries, through a methodology based on Item Response Theory (IRT), namely the Rasch Model.

On several tests, various items flagged DIF, which can compromise the cross-cultural equivalence of the different versions. However, the analysis performed does not allow to conclude that the tests are in general culturally biased, due to sample limitations which interfere with the quality of the available data.

This study alerts to the need for investigating the quality of item contents and the convenience of using both Statistical and Judgmental Analysis on test adaptations, as well as to the need for careful interpretation whenever selection decisions are based on adapted psychometric instruments. Furthermore, it shows evidence of the potential and the advantages of applying IRT techniques in Cross-Cultural Psychology research.

Keywords: Critical Reasoning Test Battery (CRTB), Cross-Cultural Equivalence, Differential Item Functioning (DIF), Item Response Theory (IRT), Test Translation/Adaptation.

Introdução

O fenómeno da globalização é já uma marca do quotidiano, onde a partilha de informação e de ideias influenciam as práticas e os procedimentos empresariais, entre outras áreas do mundo contemporâneo.

O crescimento das organizações e a necessidade de se expandirem constituem uma realidade que lhes permite estrategicamente diferenciar-se neste ambiente de globalização internacional. Assim, as práticas de recursos humanos têm a necessidade de se adaptar progressivamente a este paradigma emergente, onde a multiculturalidade é soberana e a presença de diversas nacionalidades numa mesma organização é recorrente.

Nestas práticas, inclui-se com alguma frequência a avaliação psicológica, sendo usual o recurso a técnicas psicométricas, que avaliam tanto aptidões como aspectos da personalidade. Nos processos de avaliação, devido a convenções sociais, as pessoas tendem a aceitar um certo grau de fracasso decorrente de factores aleatórios, não o vendo como injustiça. Contudo, a percepção de equidade já não se verifica quando um instrumento de medida contém itens considerados irrelevantes para um determinado processo de selecção, o que geralmente está associado à pertença a um grupo social específico, originando o enviesamento dos resultados (Rust & Golombok, 1999) e tendo por potencial consequência a discriminação de determinados grupos demográficos.

As situações supramencionadas suscitam a problemática da construção e avaliação do funcionamento de um teste numa população, nomeadamente, a sua isenção ou imunidade à influência cultural. No entanto, devido ao desenvolvimento da investigação e da prática da psicologia em diversos países, tem aumentado a difusão internacional de instrumentos de medida, tornando premente a sua adaptação para países e culturas distintas do contexto em que são originados (Prieto & Almeida, 1997).

A adaptação de instrumentos de medida resume-se, na maior parte dos casos, a uma questão linguística, o que leva os tradutores a procurarem conceitos, palavras e expressões que são cultural, psicológica e linguisticamente equivalentes (van de Vijver & Poortinga, 2005), embora colocando geralmente a tónica numa "equivalência linguística" e não numa "equivalência cultural".

A avaliação e a medição psicológica num contexto internacional devem, contudo, obedecer aos requisitos metodológicos da equivalência intercultural (Duarte & Rossier, 2008). Assim, para que qualquer comparação seja válida entre diferentes línguas e/ou grupos

culturais, todos os testes devem ser considerados equivalentes, mesmo que não se verifique uma distribuição equitativa do construto medido nos diferentes grupos, já que na comparação entre culturas é possível para um grupo a obtenção de pontuações mais altas do que para outro. Deste modo, é importante assegurar que as diferenças nos resultados observados não são devidas a falhas do teste na obtenção de resultados equivalentes, pelo que é essencial a identificação e eliminação dos itens culturalmente enviesados, promovendo a validade e precisão das medidas em contextos distintos, a fim de providenciar uma equivalência intercultural aceitável (Hambleton & Kanjee, 1995).

Neste sentido, a presente investigação não pretende contribuir para a construção da ciência psicológica, mas antes aplicar o conhecimento já disponível à resolução de problemas práticos, nomeadamente no estudo da equivalência intercultural de uma bateria de testes de aptidões, em utilização em Portugal, Reino Unido, Moçambique, Austrália e África do Sul, e envolve a avaliação da adequação deste instrumento, e do conjunto dos seus itens, nos vários países ou nas várias culturas.

Enquadramento Teórico

Avaliação Psicológica nas Organizações: uma Problemática Intercultural

A avaliação psicológica é uma prática frequente na medição das competências dos actuais e futuros colaboradores das organizações, bem como do seu potencial de desenvolvimento. Esta possibilidade de medir é encarada hoje como uma ferramenta extremamente importante, na medida em que favorece a empresa, sob o ponto de vista táctico e estratégico, culminando no aumento da sua eficácia (Bartram, 2004).

Neste âmbito, torna-se imprescindível para as organizações que recorrem aos instrumentos de medida, ou para as consultoras prestadoras de serviços nesta área, a utilização de uma multiplicidade de ferramentas de avaliação das aptidões, das competências psicomotoras e de aspectos da personalidade (Sousa, 1999). Tome-se como exemplo a prática comum de utilização destes métodos nos processos de selecção, para avaliar os atributos necessários a um desempenho eficaz de uma determinada função (Gomes et al., 2008).

Tendo como premissa as noções de globalização e a decorrente expansão das empresas e das práticas de recursos humanos, torna-se imperioso que estas últimas se generalizem, permitindo a sua implementação à escala global, o que leva a que as metodologias aplicadas num determinado país sejam transpostas para outros com diferentes línguas e culturas.

A avaliação de colaboradores neste contexto surge como uma realidade para as organizações que se expandem e adoptam as mesmas práticas internacionalmente, originando a problemática da utilização dos mesmos testes em diferentes países (Byrne et al., 2009). Assim, as empresas avaliam os actuais e potenciais colaboradores com a mesma metodologia, isto é, são comparadas pessoas com diferentes *backgrounds* nacionais, linguísticos e culturais, o que pode suscitar questões relativamente ao impacto da cultura no resultado de um teste e na sua comparabilidade.

O aumento da preocupação com a discriminação e com as políticas de promoção da igualdade de oportunidades nas organizações conduziu a um maior foco na problemática da justiça na avaliação psicológica. Por exemplo, Ryan, McFarland, Baron e Page (1999) salientam a influência dos factores nacionais e culturais nas práticas de selecção em diferentes países, aconselhando então as organizações internacionais a implementarem campanhas que promovam avaliações assentes nesta questão.

As preocupações supracitadas convergem para a necessidade de multiplicar os estudos interculturais, de forma a avaliar o impacto das potenciais diferenças entre culturas nas práticas de avaliação psicológica nas organizações (Bartram, 2004).

Construção e Adaptação de Testes em Contexto Intercultural

A construção de testes para uso em diferentes culturas gera alguma controvérsia no meio científico, pois influencia em grande escala a própria abordagem à comparação intercultural. Anastasi e Urbina (1997) referem que existem três formas de abordar a medição psicológica intercultural.

Na primeira, a escolha dos itens é feita simultaneamente nas várias culturas e a validação das medidas tem em conta um critério específico proveniente de cada uma delas, garantindo que o teste avalia equitativamente diversas culturas.

Uma segunda abordagem sugere que o teste pode ser desenvolvido numa cultura e posteriormente aplicado a participantes com diferentes *backgrounds* culturais. Neste caso, não há garantia de que um construto definido por uma determinada cultura possa ter o mesmo significado noutra. Em termos práticos, não se pode assumir que um resultado baixo num teste possa ter a mesma explicação para membros de diferentes culturas.

A terceira defende que diferentes testes podem ser desenvolvidos dentro de diferentes culturas, validados com referência a critérios locais, e utilizados apenas dentro de cada uma, invalidando qualquer possibilidade de comparação intercultural.

Por sua vez, Pike (1966, citado por Davidson, Jaccard, Triandis, Morales, & Diaz-Guerrero, 1976) apresenta um enquadramento distinto para o estudo dos fenómenos interculturais: as perspectivas *Emic* e *Etic*.

A abordagem *Emic* assenta no reconhecimento das especificidades de cada cultura, salientando que devem ser compreendidas no seio das suas particularidades. Em oposição, a abordagem *Etic* apoia-se na universalidade ou transversalidade cultural das leis, admitindo que os construtos são conceptualmente equivalentes em todas as culturas.

Estas perspectivas têm evidente impacto na investigação intercultural, na medida em que quem adopta a perspectiva *Emic* encontra benefícios na obtenção de descrições culturais muito específicas, embora arrisque comprometer a contemplação de factores comuns, e quem adopta a perspectiva *Etic* aposta na categorização universal dos fenómenos, o que possibilita comparações culturais, sacrificando contudo o objecto ou a temática central do estudo (Prieto & Almeida, 1997).

Esta dificuldade de obtenção de descrições de um fenómeno, concomitantemente específico a uma cultura e comparável a outras, foi descrito como o "dilema *Etic-Emic*" (Berry, 1969, citado por Davidson et al., 1976). Vários autores tentaram ultrapassar esta discrepância entre a especificidade (*Emic*) e universalidade (*Etic*) tais como Davidson e colaboradores (1976) e Prieto e Almeida (1997), destacando-se os primeiros por apresentarem duas soluções.

A primeira, e mais frequente, é a abordagem "*Pseudoetic*", onde as medidas *Emic* são assumidas como *Etic*, por exemplo, um instrumento composto por itens que reflectem a cultura ocidental é traduzido e utilizado noutras culturas, sem considerar que as diferenças nas medidas significam diferenças culturais.

A outra abordagem resulta da combinação *Etic-Emic*. Neste caso o investigador identifica um construto *Etic* (que tem um estatuto universal) e posteriormente define caminhos *Emic* para medir esse construto, elaborá-lo e validá-lo. Por outras palavras, são desenvolvidos métodos *Emic* para a avaliação de construtos *Etic*. Neste sentido, Prieto e Almeida (1997) propõem um método para conseguir a equivalência transcultural, que pode partir do recurso a instrumentos constituídos simultaneamente por itens específicos de uma dada cultura e por itens comuns a outras culturas.

Este interesse cada vez maior pela diversidade cultural, ao nível da investigação e da medição psicológica para comparação de grupos culturais intra e inter-países, levou Byrne e colaboradores (2009) a afirmarem que esta temática se tornou uma moda, podendo daí advir diversos problemas. Primeiro, devido ao crescente número de investigadores que se

interessam por esta área, carecendo de suficiente formação em psicologia intercultural. Segundo, porque estas investigações envolvem o uso de testes e outros dados, que primeiro requerem adaptação, para poderem ser utilizados noutra cultura. E por último, porque estas questões exigem um profundo conhecimento do tema, bem como de *design* de investigação e de metodologias de análise de dados, algo que nem todos os investigadores dominam.

Tendo em conta os factores enumerados, e acrescendo a estes a ampliação das trocas internacionais e a multiplicação de exames para conceder credenciais internacionais (equivalências/créditos), Hambleton (2005) afirma que é expectável um crescimento do número de adaptações interculturais de testes. A literatura enumera cinco razões que justificam este facto (Hambleton & Kanjee, 1995; Hambleton & Patsula, 1999): (1) o processo envolve menos custos e é mais célere que o desenvolvimento de um novo instrumento; (2) a produção de testes equivalentes para comparação entre grupos étnicos e culturais verifica-se ser mais fácil e eficaz; (3) a produção de um novo instrumento não necessita de um conhecimento tão aprofundado em termos conceptuais e metodológicos; (4) o conhecimento da eficácia do instrumento confere uma sensação de confiança e segurança, principalmente quando o teste original é conhecido; (5) a maior equidade dos respondentes que realizam o teste, na medida em que lhes é permitido serem avaliados num idioma à sua escolha.

A adaptação de testes inclui todo um conjunto de actividades, como a própria ponderação da possibilidade do teste medir o mesmo construto em diferentes línguas e culturas, a selecção de tradutores, a adaptação em si ou a verificação posterior da sua equivalência (Hambleton, 2005). Alguns investigadores consideram mesmo uma percentagem considerável dos estudos nesta área como sendo infrutíferos, ou até mesmo inválidos, devido a uma adaptação inadequada dos testes (van de Vijver & Poortinga, 2005).

Em 1992 a *International Test Commission* (*ITC*) iniciou um projecto de preparação de directrizes para orientar e regulamentar a tradução e adaptação de testes e instrumentos de avaliação psicológica, e estabelecer a equivalência de resultados em diferentes línguas e/ou culturas (*ITC*, 2010). Mais tarde, em 1994, foi publicado um relatório sobre o progresso destas directrizes e, em 1999, um estudo sobre a sua aplicação prática (Gregoire & Hambleton, 2009). Estas directrizes – *Guidelines for Translating and Adapting Tests* – têm a sua aplicação em dois contextos: na tradução/adaptação de testes ou instrumentos de avaliação já existentes ou no desenvolvimento de novos testes e instrumentos para utilização internacional (*ITC*, 2010), devendo ser consideradas como uma boa prática e uma referência que responsabiliza cada utilizador por seleccionar o melhor procedimento disponível para as

respeitar, não se limitando a encará-las como uma mera lista de procedimentos psicométricos (Gregoire & Hambleton, 2009).

No primeiro contexto destas directrizes, o objectivo é produzir uma versão de um teste ou instrumento de avaliação comparável ao original, no que diz respeito às suas características psicométricas. O segundo contexto reporta-se a situações específicas em que ocorre o desenvolvimento de testes e instrumentos com o objectivo de fazer comparações internacionais. Neste caso, as versões utilizadas nas diferentes línguas ou contextos culturais são desenvolvidas simultaneamente, o que se torna vantajoso, pois não há a necessidade assegurar previamente um conjunto de qualidades psicométricas (*ITC*, 2010).

Além das *Guidelines for Translating and Adapting Tests*, também uma outra publicação, os *Standards for Educational and Psychological Testing* foi desenvolvida com o propósito de estabelecer princípios fundamentais dirigidos aos psicólogos ou a outros especialistas relativamente à selecção, desenvolvimento e aplicação de testes psicológicos. Alguns destes *standards* alertam os utilizadores para as fontes de erro ou de enviesamento que afectam a adaptação de um teste a uma cultura e/ou língua diferente da original. Estas fontes de erro podem ser divididas em três categorias: (a) diferenças culturais/linguísticas; (b) problemas técnicos, de construção ou de método e (c) problemas de equivalência na interpretação dos resultados. Caso estas noções não sejam contempladas, pode construir-se um teste não equivalente nas duas línguas ou culturas, o qual conduz a interpretações erróneas e origina conclusões infundadas, relativamente aos grupos envolvidos. (Hambleton, 2005).

Em suma, o paradigma de multiculturalidade emergente alerta para a necessidade de utilização de diversas versões dos testes e instrumentos de avaliação em diferentes línguas, para serem utilizados num único contexto nacional (*ITC*, 2010).

Noções de Equivalência Intercultural e de Enviesamento Cultural

No contexto da adaptação de testes, são dois os conceitos essenciais que devem ser considerados quando se avalia a adequação de um instrumento de medida em diferentes culturas: Equivalência e Enviesamento.

Noção de Equivalência Intercultural

O conceito de equivalência está relacionado com a comparabilidade dos resultados obtidos em diferentes contextos culturais.

van de Vijver e Leung (1997) referem primeiramente que quando um instrumento mede diferentes construtos em duas culturas, nenhuma comparação pode ser efectuada, ou

seja, não pode ser realizada a transposição dos resultados obtidos em diferentes grupos culturais (por exemplo, "comparar maçãs com laranjas"). Neste caso, ocorre uma situação de Não Equivalência do Construto, coincidente com a perspectiva *Emic* (van de Vijver & Tanzer, 2004). No entanto, quando o construto pode ser comparado, existem três níveis de equivalência psicométrica que são crescentemente mais complexos.

O primeiro diz respeito à Equivalência de Construto (também denominada "equivalência estrutural" ou "funcional") que ocorre quando o significado e a estrutura dimensional de um construto são idênticos em diferentes grupos culturais. Quando existe este tipo de equivalência, o mesmo construto é medido nas diferentes culturas, embora eventualmente através de diferentes operacionalizações (van de Vijver & Leung, 1997). Exemplificando com o conceito de culpa: se em diferentes culturas forem identificadas situações que conduzem à culpa, e a partir destas se desenvolver um instrumento, pode-se medir a culpa em cada grupo cultural, no entanto os resultados não podem ser comparados. Encontra-se neste caso implícita uma validade universal subjacente ao construto psicológico, concordante com a perspectiva *Etic* (van de Vijver & Tanzer, 2004).

O segundo nível é a Equivalência da Unidade de Medida e sucede quando as duas métricas têm a mesma unidade de medida mas tomam por referência diferentes origens (é o caso as escalas de temperatura Kelvin e Celsius). Por outras palavras, perante a existência de uma diferença numa escala, esta verifica-se simultaneamente na outra, tornando-as equiparáveis (van de Vijver & Tanzer, 2004). Este tipo de equivalência pressupõe resultados intervalares ou de razão (a mesma unidade de medida em todas as culturas) (van de Vijver & Leung, 1997).

O último nível, mais complexo, é a Equivalência do Resultado Total (*Scalar*), que pode ser alcançado quando a origem da escala é igual para diferentes grupos culturais viabilizando a possibilidade de comparações directas, tanto ao nível individual como intercultural (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004). A forma mais simples de compreender este tipo de equivalência é tomar como exemplo o comprimento (entre centímetros e polegadas) ou o peso (entre quilogramas e onças).

Noção de Enviesamento Cultural

Para duas versões serem consideradas equivalentes, não podem existir enviesamentos (van de Vijver & Poortinga, 1997). Este é um termo genérico que indica a falta de correspondência entre os resultados observados em diferentes culturas e o domínio da generalização (traço ou aptidão que o teste mede), sendo importante salientar que resulta da

presença de factores de ruído que põem em causa a validade das comparações interculturais e que podem ter diferentes origens: traduções erradas, conteúdos dos itens inadequados, falta de estandardização na aplicação, entre outros (van de Vijver & Leung, 1997).

No presente estudo, distinguem-se três tipos de enviesamento – Construto, Método e Item – que podem afectar os construtos teóricos, os procedimentos de pesquisa ou a análise de resultados, e que têm diferentes origens, o que implica a adopção de distintos procedimentos para os detectar e ultrapassar (van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1997; van de Vijver & Tanzer, 2004).

Enviesamento do Construto: ocorre quando a definição de um construto não é igual em duas culturas diferentes. van de Vijver e Poortinga (1997) referem que este enviesamento surge quando: (1) apenas uma parte da definição do construto é comum às culturas em causa; (2) há diferença nos comportamentos associados ao construto (apenas uma das culturas manifesta o comportamento); (3) apenas uma pequena amostra do comportamento é relevante (e.g., instrumentos pequenos); (4) a cobertura dos aspectos/facetas relevantes do construto é incompleta. van de Vijver e Tanzer (2004) afirmam que, para ultrapassar este tipo de enviesamento, os investigadores podem fazer um descentramento cultural (*descentering*), e construir um mesmo instrumento em diversas culturas simultaneamente, ou utilizar uma abordagem convergente, na qual este se desenvolve de forma independente dentro de uma cultura, para posteriormente ser aplicado noutras.

Enviesamento do Método: inclui todas as variáveis de ruído relacionadas com os factores do método, descritos nas secções empíricas dos artigos científicos, e pode ser de três tipos: amostra, aplicação e instrumento (van de Vijver & Tanzer, 2004).

O enviesamento ou incomparabilidade das amostras ocorre quando estas diferem em características relevantes para além do construto a ser estudado, por exemplo, quando se fazem comparações entre culturas "remotas" (isto é, que diferem em muitos aspectos) pode ser impossível estabelecer uma correspondência entre elas.

O enviesamento de aplicação pode surgir quando há diferenças físicas, técnicas (e.g. câmaras de vídeo) e/ou sociais (e.g. tamanho do grupo) nas condições de aplicação, podendo também decorrer de ambiguidade nas instruções dadas aos participantes e/ou aplicadores, dos efeitos induzidos pelo aplicador (e.g. Efeito de Halo), assim como de problemas de comunicação entre este e o respondente (e.g. problemas de interpretação ou inclusão de temas tabu).

Por fim, o enviesamento do instrumento refere-se às questões de familiaridade com o tipo de material estímulo (e.g. itens), e com os estilos (e.g. desejabilidade social, aquiescência) e procedimentos de resposta.

Para identificar e colmatar os efeitos do Enviesamento de Método, van de Vijver e Tanzer (2004) identificam as seguintes estratégias: (1) Formação extensiva dos aplicadores; (2) Manuais detalhados de aplicação, cotação e interpretação; (3) Instruções detalhadas (e.g. com número suficiente de exemplos e/ou exercícios); (4) Utilização de variáveis do contexto do participante (e.g. *background* académico/formativo – ao obter informação sobre variáveis estranhas, torna-se possível verificar estatisticamente a sua influencia); (5) Utilização de informação colateral (e.g. comportamento na aplicação do teste); (6) Avaliação dos estilos de resposta; e (7) Realização de estudos teste-reteste.

No âmbito dos Enviesamentos do Construto e do Método, van de Vijver e Tanzer (2004) referem a existência de estratégias que permitem ultrapassá-los em conjunto, as quais passam pela utilização de especialistas na cultura e língua local, pelo recurso a amostras bilingues, assim como pela comparação intercultural de redes nomológicas de estruturas de relações entre construtos (e.g. estudos de validação convergente/discriminante).

Enviesamento do Item: decorre de diferentes fontes, sendo que as mais comuns consistem na tradução imprópria do item, na ambiguidade do item original, na pouca familiaridade/adequação do conteúdo do item a uma determinada cultura, nos factores de ruído (e.g. o item evoca outros traços e/ou aptidões) e nas influências específicas da cultura como o uso de um termo de significado específico ou de algumas particularidades da linguagem coloquial (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004). Enquanto os Enviesamentos do Construto ou do Método dizem respeito a características mais gerais do instrumento, o Enviesamento do Item põe em causa a validade, devido à dependência da qualidade dos itens (van de Vijver & Poortinga, 2005). Os enviesamentos associados à tradução do item são geralmente discutidos ao nível da linguística e representam um caso especial de equivalência da medida, quando um instrumento original é traduzido para outra língua (Byrne et al., 2009).

A evolução da psicometria e em particular das técnicas de detecção do funcionamento de itens anómalos ou desajustados levaram à substituição do tradicional termo "Enviesamento do Item" pelo termo *Differential Item Functioning* (*DIF*) ou, em português, Funcionamento Diferencial do Item (van de Vijver & Poortinga, 2005).

O *DIF* permite identificar os itens em que pessoas com o mesmo nível de competência num determinado construto (operacionalizado como o resultado total) não têm o mesmo resultado expectável num determinado item (operacionalizado como o resultado médio do item) (van de Vijver & Poortinga, 2005), ou seja, pessoas do mesmo nível de competência provenientes de diferentes grupos culturais têm diferentes probabilidades de sucesso num mesmo item, sendo que o mesmo nível de competência significa igual "quantidade" do construto que se pretende medir (Anastasi & Urbina, 1997). Ainda neste contexto, Mellenberg (1982, citado por van de Vijver & Poortinga, 2005) fez uma distinção entre enviesamento uniforme e não uniforme, sendo que o enviesamento de um item é uniforme se a diferença no nível de desempenho for relativamente constante ao longo dos níveis de "aptidão" e não uniforme quando essa condição não se verifica.

Relativamente às estratégias para identificar e colmatar o Enviesamento do Item, van de Vijver e Tanzer (2004) propõem para a detecção de enviesamento: (1) o Método dos Juízes (*Judgemental Analysis*) (e.g. análise linguística e psicológica); (2) os Métodos Estatísticos (e.g. análise do *DIF*); (3) a Análise de erros e distractores; e (4) a criação de "itens suplentes" que representem uma boa medida do construto, e possam servir como substitutos aos itens que tenham funcionamento diferencial.

Vários Métodos Estatísticos têm sido empregues no estudo do *DIF*, como o Método Delta Plot (Angoff, 1972, 1993, citado por Sireci, Patsula, & Hambleton, 2005) ou o Método Mantel-Haenszel (Holland & Thayer, 1988), porém, com o aumento da utilização dos computadores, os procedimentos baseados na Teoria da Resposta ao Item (TRI) apresentaram-se como mais promissores (Anastasi & Urbina, 1997; Sireci, Patsula, & Hambleton, 2005), e têm vindo a ser crescentemente aplicados em investigação intercultural.

Relacionado com o *DIF* existe na literatura outro conceito que merece destaque: Funcionamento Diferencial do Teste (*Differential Test Functioning – DTF*). Contrariamente ao *DIF*, que se foca na análise do item, o *DTF* centra-se na compreensão holística do funcionamento do teste e se este é, na sua globalidade, susceptível à discriminação entre grupos (Drasgow & Probst, 2005).

Implicações do Enviesamento Intercultural para a Equivalência Intercultural

O enviesamento e a equivalência são dois conceitos indissociáveis. O primeiro pode ou não reduzir o nível do segundo, determinando assim o tipo de comparação viável, entre diferentes culturas (van de Vijver & Leung, 1997).

Na Equivalência do Construto, apenas o Enviesamento do Construto tem influência; quanto à Equivalência do Resultado Total, todos os tipos de enviesamento a condicionam; a Equivalência da Unidade de Medida é influenciada pelo Enviesamento do Construto, pelo Enviesamento do Método e pelo Enviesamento do Item não uniforme. Neste último caso, o Enviesamento do Item uniforme não influencia a equivalência, uma vez que as diferenças existentes são constantes ao longo dos níveis de competência no construto.

Em suma, as comparações directas de resultados requerem um nível mais elevado de equivalência (o que pressupõe um menor enviesamento) do que as restantes comparações. Contudo, o Enviesamento do Construto é o desafio principal à comparabilidade dos resultados, pois caso não seja eliminado, introduz uma forma de não equivalência que exclui qualquer possibilidade de comparação intercultural (van de Vijver & Poortinga, 2005).

Vantagens da Teoria da Resposta ao Item para o estudo intercultural de instrumentos de medida

A construção e análise metrológica de testes obedece a modelos que condicionam cada passo, sob o ponto de vista teórico e estatístico (Muñiz, 2010). Neste sentido, há duas grandes teorias (ou modelos de medida) que se destacam, pela sua ampla divulgação e aplicabilidade: a Teoria Clássica dos Testes (TCT) e Teoria da Resposta ao Item (TRI).

A TCT, também conhecida pela Teoria do Resultado-Verdadeiro (Anastasi & Urbina, 1997; Muñiz, 2010), inicialmente proposta por Spearman (1904), tornou-se o modelo de medida predilecto dos investigadores no desenvolvimento de instrumentos de medida, devido à sua facilidade de aplicação em diferentes situações de medição psicológica (Hambleton & Jones, 1993). No modelo TCT, a variável dependente é o resultado total da prestação de um examinado num teste, e as variáveis independentes são o resultado verdadeiro no traço avaliado e o erro de medida, sendo que o primeiro consiste na pontuação que alguém obteria, em média, se realizasse infinitas vezes o teste. Contudo, uma vez que esta condição é impossível, os investigadores apenas podem considerar o resultado observado, ou pontuação empírica obtida num teste, à qual está obviamente associado um erro que pode ter diversas origens, tais como a própria pessoa, o próprio teste ou o contexto (Dickes, Tournois, Flieller, & Kop, 1994; Fan, 1998; Muñiz, 2010; Urbina, 2004) e que pode ser estimado através dos métodos de estudo da precisão.

Ao nível do item, o modelo TCT é relativamente simples, uma vez que não é necessário evocar um modelo teórico complexo para relacionar a capacidade de um examinado responder acertadamente a cada item. Neste sentido, a TCT considera uma *pool* de

examinados e testa empiricamente a taxa de sucesso num item, sendo que esta é tomada como referência ao definir o seu nível de dificuldade (Fan, 1998).

Apesar de existirem poucas dúvidas relativamente à sua eficácia, e de um número elevado de testes serem construídos à luz do modelo TCT, várias limitações têm sido apontadas com o propósito de o desacreditar, muito designadamente, a dependência circular dos testes (Fan, 1998), isto é, o facto de a estatística de um respondente (i.e., resultado observado) ser dependente da amostra de itens com que foi examinado, ao mesmo tempo que as estatísticas de um item (i.e., dificuldade e discriminação) são dependentes da amostra de respondentes em que foram obtidas. A esta limitação acresce o facto de no modelo TCT o resultado verdadeiro se aplicar apenas aos itens de um teste específico, ou de um teste com propriedades equivalentes, dificultando a comparação entre resultados.

Como resposta a estas limitações, foi proposto um novo modelo de medida que vem demonstrar ter vantagens sobre o modelo clássico: a Teoria da Resposta ao Item (TRI) (*Item Response Theory*, *IRT*), também conhecida como Modelo de Traço Latente (Moreira, 2004). A TRI, como o próprio nome indica, centra-se ao nível do item, em oposição ao foco no resultado total do teste, presente na TCT. Este modelo baseia-se na determinação da probabilidade de cada resposta em função do nível de competência de cada respondente no traço latente e de parâmetros dos itens (dificuldade, discriminação ou *guessing*), sendo que o padrão de respostas de um respondente a um conjunto de itens proporciona uma base para estimar o seu nível no traço latente (Embretson & Reise, 2000).

Os modelos da TRI baseiam-se em três postulados. O primeiro é a independência local, e afirma que a probabilidade de resposta a um item não depende das respostas aos outros itens. O segundo assume que o resultado total na escala depende exclusivamente da característica ou atributo (do construto) que está a medir, conferindo unidimensionalidade a essa escala. Por fim, o terceiro estabelece que as Curvas Características dos Itens (CCI) têm uma forma específica, correspondente ao gráfico da função matemática que relaciona a probabilidade de uma determinada resposta a cada item com o nível de competência do sujeito na dimensão a medir (Anastasi & Urbina, 1997).

Os modelos da TRI variam consoante a existência de um, dois ou três parâmetros dos itens. Os mais aplicados são os Modelos de um parâmetro (sendo o mais conhecido o Modelo de Rasch, aplicável a itens de cotação dicotómica), onde é avaliada a dificuldade de cada item como o ponto ou nível da dimensão na qual a probabilidade de acerto é .50. Nos Modelos de dois parâmetros, para além da dificuldade, é também contemplada a discriminação dos itens, representada pelo declive das CCI. E os Modelos dos três parâmetros, para além dos já

mencionados, contemplam também o *guessing*, representado pela assímptota inferior da curva e que diz respeito à probabilidade de acerto ao acaso - por exemplo, em casos de escolha múltipla com cinco opções, existe no mínimo uma probabilidade de acerto de .20 (Anastasi & Urbina, 1997; Bond & Fox, 1998; Embretson & Reise, 2000; Fan, 1998).

Os métodos TRI oferecem a vantagem de poder assumir-se como invariantes, pelo que proporcionam uma escala uniforme que pode ser utilizada em diferentes grupos, salvaguardando que os dados se ajustam ao modelo e que alguns pressupostos do modelo sejam cumpridos. O nível de competência nas aptidões ou nos traços estimados é independente do conjunto de itens particulares aplicados a uma determinada amostra, sendo também os parâmetros dos itens independentes da amostra em que foram obtidos (invariância específica), o que permite ultrapassar a supra referida circularidade na TCT. Por último, quando na aplicação de um teste se adoptam procedimentos adaptativos, os erros estimados resultantes são independentes do conjunto de itens aplicados a um determinado indivíduo (Urbina, 2004).

Os vários métodos de identificação de *DIF* na TCT são dependentes da amostra, sendo necessário partir da suposição de que resultados observados correspondem ao construto que se pretende medir (Scheuneman & Bleinstein, 1989). Um dos exemplos mais referidos é o Delta Plot (Angoff, 1972, 1993, cit. por Sireci et al., 2005), que calcula a proporção de respostas correctas nos grupos em causa, colocando-as num gráfico, sendo que os itens têm a mesma dificuldade quando na referida representação gráfica surge uma linha com um ângulo de 45º. Por outro lado, os métodos de detecção de *DIF* que recorrem à TRI, partem dos parâmetros das CCI, que são independentes relativamente à amostra (Scheuneman & Bleinstein, 1989). Assim, se os parâmetros são estimados separadamente para cada grupo, e se o comportamento do item é semelhante nos dois grupos, então as CCI serão iguais. Uma outra vantagem é que a comparação entre grupos é realizada a todo o nível de competência, podendo inclusive identificar através da CCI, os níveis de competência em que ocorre *DIF*.

Problema, objectivos e hipótese

A presente investigação tem como objectivo contribuir para o aprofundamento da temática da equivalência intercultural, recorrendo às vantagens da Teoria da Resposta ao Item para este tipo de estudos. Com a quantidade de adaptações que têm sido realizadas para um grande número de instrumentos, verifica-se ser pertinente averiguar se os testes que foram adaptados, e que estão a ser actualmente utilizados em diversos países, são culturalmente equivalentes. Como a questão da equivalência de um teste depende obviamente do

instrumento em questão, não será possível esboçar um problema transversal à literatura, sendo para tal necessário circunscrever o instrumento específico a tratar. No presente caso estuda-se uma bateria de três testes (Verbal, Numérico e Diagramático) – Bateria de Raciocínio Crítico (*CRTB*) – em utilização em Portugal, Reino Unido, Moçambique, Austrália e África do Sul.

Tendo como premissa que os construtores de testes nos processos de adaptação das versões em estudo seguem um conjunto de etapas que visam promover a sua equivalência, e consequentemente reduzir a probabilidade de enviesamento, o presente estudo assenta numa única hipótese, aplicável a todos os testes da bateria:

Hipótese: As versões adaptadas dos testes são equivalentes às versões originais.

Método

Participantes

Para efeitos do presente estudo obtiveram-se 4946 respostas em contextos de avaliação psicológica, em diversas organizações nacionais e internacionais. Contudo, importa salientar que este valor não corresponde ao número total de participantes, uma vez que há sujeitos que responderam cumulativamente aos três testes em estudo. Deste modo, torna-se relevante descrever a presente amostra por tipo de teste realizado.

Relativamente ao teste Verbal, auferiram-se 645 respostas, das quais 23.6% são de Portugal, 4.9% de Reino Unido e 71.5% de Moçambique. Neste grupo, 95.1% responderam ao teste na língua portuguesa e 4.9% na língua inglesa.

Quanto à idade e ao género dos respondentes, devido a inultrapassáveis constrangimentos na obtenção dos dados, apenas se podem descrever as amostras provenientes de Portugal e de Moçambique. Assim, em Portugal, a faixa etária varia entre os 19 e os 55 anos de idade (M=29.53; DP=6.99), sendo que 52.2% são mulheres. Já em Moçambique, as idades variam entre os 19 e os 44 anos (M=26.72; DP=3.83) sendo 60.4% dos respondentes mulheres.

Os participantes que responderam ao teste Numérico eram de Portugal, Reino Unido e Moçambique, sendo as respectivas percentagens de 23.6%, 5.4% e 71.0%. Neste teste, 94.6% responderam na língua portuguesa e 5.4% na língua inglesa.

À semelhança do teste Verbal, apenas se recuperou a idade e o género dos participantes portugueses e moçambicanos. Em Portugal, os respondentes ao teste Numérico também responderam ao teste Verbal, pelo que a média de idades e percentagens de géneros

são as mesmas. Em Moçambique, a faixa etária varia entre os 19 e os 44 anos (M=26.72; DP=3.83), sendo 60.3% mulheres.

Por último, no teste Diagramático, obtiveram-se 3767 respostas, das quais 3.8% são de Portugal, 2.6% de Reino Unido, 11.1% de Moçambique, 74.8% da Austrália e 7.7% da África do Sul.

Uma vez mais, apenas se obtiveram os dados demográficos de Portugal e Moçambique. Relativamente ao primeiro grupo, as idades variam entre os 19 e os 49 anos (M=25.80; DP=4.20) e 42.3% são mulheres. Em Moçambique, 60.4% dos respondentes são mulheres e a faixa etária situa-se entre os 19 e os 44 anos (M=26.73; DP=3.85).

O referido constrangimento, ao nível da obtenção de dados demográficos, estendeu-se também à obtenção de informações relativas à escolaridade e à profissão dos participantes, por não estarem incluídas nas diversas bases de dados.

Instrumento

A *Critical Reasoning Test Battery* (*CRTB*) ou, em Português, Bateria de Raciocínio Crítico, foi construída originalmente em 1982 no Reino Unido e utilizada para fins de selecção ou como técnica auxiliar de aconselhamento profissional. Em 1991, a revisão desta bateria consolidou a sua utilização no ambiente profissional, proporcionando a medida das seguintes aptidões: raciocínio crítico verbal, raciocínio crítico numérico e raciocínio crítico diagramático. Com efeito, esta bateria destina-se a Chefias Directas e Intermédias, Técnicos Especializados, Administrativos ou Comerciais, Bacharéis e Licenciados com experiência profissional pouco significativa (1991a).

De forma a compreender melhor a sua composição desta bateria, apresenta-se de seguida uma descrição sucinta de cada teste.

O teste Verbal pretende avaliar a capacidade de compreender e analisar informação escrita, de forma a obter conclusões lógicas. Neste teste, são apresentados catorze textos seguidos de quatro afirmações, sendo solicitado ao respondente que analise cada afirmação e responda, com base no texto, se a afirmação é verdadeira, falsa ou se não se pode afirmar se é verdadeira ou falsa sem informações adicionais. Na versão inglesa este teste é constituído por 60 itens e na portuguesa por 56, partilhando ambas as versões 26 itens. O tempo máximo de realização é respectivamente de 30 e 25 minutos. O teste aplicado em Moçambique é uma versão reduzida da versão portuguesa, em que existem apenas três (em alguns casos, duas) afirmações por texto, perfazendo um total de 30 itens com um tempo limite de 20 minutos.

O teste Numérico avalia a capacidade de raciocinar com dados numéricos apresentados em quadros estatísticos. O teste é constituído por oito gráficos e tabelas que contêm diferentes tipos de conteúdos, sendo apresentadas cinco questões relativas a cada um destes quadros, distribuídos uniformemente pelo teste. Para cada questão, o respondente tem de ler correctamente o quadro, seleccionar os dados e efectuar os cálculos necessários, de forma a escolher a opção de resposta correcta. Em cada item, são apresentadas cinco alternativas, onde por vezes se inclui a opção "Não se pode saber". Nas versões portuguesa e inglesa, o teste é composto por 40 itens, e tem um tempo máximo de realização de 30 minutos. A versão moçambicana é novamente reduzida, com 18 itens cujo tempo de realização é de 20 minutos.

Por último, o teste Diagramático avalia a capacidade de raciocinar e compreender a lógica entre sequências, constituídas por material simbólico. Neste teste não há conteúdo verbal ou numérico e em cada item é apresentada uma sequência de cinco diagramas, tendo o respondente de seleccionar o seguinte, a partir de um conjunto de cinco opções. Em todas as versões o teste é composto por 40 itens e tem 20 minutos de tempo limite de realização, exceptuando a versão moçambicana, que é constituída por 30 itens e tem 15 minutos de realização.

Estes testes são respondidos no formato papel e lápis, e os participantes registam a opção escolhida numa folha de respostas. Previamente à realização do teste, existe um período de instruções e exemplos, para que os respondentes se familiarizem com a tarefa a realizar.

No que concerne aos índices metrológicos destes testes, foram conduzidos estudos de validade que concluem que os itens presentes reflectem o tipo de raciocínio geralmente exigido para o nível a que o teste se propõe medir (SHL, 1991a). Adicionalmente, os índices de consistência interna deste instrumento (*alfa* de Cronbach) são para o teste Verbal, Numérico e Diagramático respectivamente de .90, .86 e .81 na versão inglesa (SHL, 1991b) e de .89, .89 e .86 na versão portuguesa (SHL, 1991a). Os dados mencionados justificam a escolha desta bateria para a presente investigação.

Procedimento

Os dados foram disponibilizados por uma empresa de consultoria portuguesa que utiliza esta bateria e que estabeleceu o contacto com uma consultora pertencente ao mesmo grupo organizacional e com empresas estrangeiras clientes. As folhas de respostas moçambicanas foram enviadas para Portugal e os dados ingleses, australianos e sul-africanos foram obtidos por via electrónica.

Nas várias versões em estudo a ordenação dos itens não é a mesma, pelo que se tornou necessário proceder ao estabelecimento de correspondência entre os mesmos, tomando como referência a versão portuguesa.

Após a correspondência entre os itens, procedeu-se à análise estatística dos dados, pela aplicação de técnicas de Análise TRI (Modelo de Rasch), através do programa WINSTEPS 3.70.0 (Linacre, 2010). O facto de o teste não ter um formato de resposta obrigatória permite a existência de respostas omissas, sendo que nesta análise estas não foram consideradas como respostas erradas (cotação "zero"), para que houvesse possibilidade de distinção entre erro e ausência de resposta.

Para cada versão dos testes, foi primeiro realizada uma análise separada, a fim de estudar as medidas proporcionadas pelos itens respectivos e, de seguida, procedeu-se a análises conjuntas para estabelecer comparações entre os dados obtidos nos diversos países e detectar a eventual presença de Funcionamento Diferencial dos Itens (*Differential Item Functioning – DIF*).

Métodos de análise de resultados

A metodologia TRI, em particular o Modelo de Rasch, só é útil se as medidas obtidas se ajustarem ao modelo, sendo este visto como referência de uma medida ideal (Bond & Fox, 1998). Neste sentido, o primeiro tipo de análise a efectuar pretende averiguar em que medida os resultados obtidos com um dado instrumento, e a partir de uma determinada amostra, se ajustam ao modelo, e qual o grau desse ajustamento. Contudo, importa salientar que é a medida que se ajusta ao Modelo de Rasch, e não o oposto pelo que, em caso de desajuste, se torna necessária uma revisão do método da medida, nunca do modelo.

O Modelo de Rasch traduz-se numa descrição matemática de variáveis psicológicas ou sociais e recorre à análise da estatística Qui-Quadrado para averiguar se os dados empíricos se ajustam ao modelo, nomeadamente através dos índices *infit* e *outfit* (Bond & Fox, 1998).

O índice *infit* (*inlier-pattern-sensitive fit*) é obtido através do Qui-Quadrado, com cada observação ponderada pela respectiva informação estatística (variância do modelo). É mais sensível a padrões inesperados de observações das pessoas em itens cuja dificuldade se situa no seu nível de atributo, ou de observação dos itens em pessoas com competência situada ao seu nível de dificuldade (Linacre, 2010). A interpretação destes índices podem variar consoante o tipo de teste (Bond & Fox, 1998) contudo, para que as informações sejam produtivas para a medição, considera-se que há ajustamento quando os índices de *infit* se situam entre .5 e 1.5 (Linacre, 2010).

Por outro lado, o índice *outfit* (*outlier-pattern-sensitive fit*), também baseado na informação estatística do Qui-Quadrado, é mais sensível a observações inesperadas de pessoas em itens muito fáceis ou muito difíceis para o seu nível de atributo, ou de itens em pessoas com níveis de competência muito baixos ou muito altos para o seu nível de dificuldade. À semelhança do índice *infit*, quando há ajustamento, os seus valores devem situar-se entre .5 e 1.5 (Linacre, 2010).

O valor esperado dos índices *infit* e *outfit*, em caso de ajustamento perfeito é de 1, que significa 100% de compatibilidade entre os padrões de resposta observados e os padrões de resposta esperados de acordo com o Modelo de Rasch (Bond & Fox, 1998).

Para a detecção de *DIF*, são geralmente constituídos dois grupos, denominados tipicamente por grupo focal e por grupo de referência. Dependente do contexto de investigação, estes grupos podem diferir na média e desvio-padrão dos seus resultados na variável latente, contudo importa referir que diferenças nestes valores não significam necessariamente que haja *DIF*, tornando crucial a realização de análises mais aprofundadas (Embretson & Reise, 2000).

À luz da metodologia do traço latente, existem diversos procedimentos para a detecção de *DIF*, como por exemplo o de Lord (1980), que refere existir *DIF* quando as CCI diferem no grupo focal e no grupo de referência, ou seja, quando são necessários diferentes parâmetros para descrever o funcionamento do item em cada grupo.

Porém, Linacre (2010) refere que o método de Lord aparentemente sobredetecta a presença de *DIF*, e apresenta outro procedimento para a sua identificação. Primeiro, reúnem-se todos os dados da amostra e produzem-se valores de ancoragem para o nível de competência dos sujeitos, para posteriormente determinar o nível de dificuldade de cada item. O passo final consiste em comparar a diferença entre os níveis de dificuldade do item nos dois grupos, recorrendo ao teste *t-Student* para verificar a sua significância estatística. Este autor refere então que os itens têm a presença de *DIF* quando $t > \pm 2.00$ ($p < .05$).

Resultados

1. Análise do ajustamento ao Modelo de Rasch

A Tabela 1 apresenta os índices de ajustamento ao modelo (*infit* e *outfit*) para cada teste em estudo. Analisando os valores de ajustamento do teste Verbal através do *infit*, pode-se verificar que as três versões têm valores médios *MNSQ* (*mean square*) entre .98 e 1.00, significando que os itens do teste Verbal, de uma forma geral, ajustam-se ao modelo.

Tabela 1

Análise de Rasch: índices de ajustamento ao modelo

Teste	Versão (nº de itens analisados)		*Infit*						*Outfit*					
			MNSQ (*Mean Square*)				>1.5	>2.0	*MNSQ* (*Mean Square*)				>1.5	>2.0
			Média	dp	Mín	Máx	F (%)	F (%)	Média	dp	Mín	Máx	F (%)	F (%)
Verbal	Portugal (56)	Itens	1.00	.06	.88	1.26	0 (0.0)	0 (0.0)	.99	.19	.57	1.81	1 (1.8)	0 (0.0)
		Sujeitos	.99	.16	.59	1.52	1 (0.7)	0 (0.0)	.99	.41	.11	3.50	10 (7.2)	2 (1.4)
	Moçambique (30)	Itens	1.00	.08	.89	1.21	0 (0.0)	0 (0.0)	1.00	.12	.82	1.31	0 (0.0)	0 (0.0)
		Sujeitos	1.00	.18	.57	1.57	3 (0.7)	0 (0.0)	1.00	.28	.45	2.55	19 (1.3)	3 (0.7)
	Reino Unido (60)	Itens	.98	.18	.64	1.39	0 (0.0)	0 (0.0)	.95	.38	.22	1.81	4 (7.0)	0 (0.0)
		Sujeitos	.99	.13	.74	1.35	0 (0.0)	0 (0.0)	.97	.31	.43	1.85	2 (6.0)	0 (0.0)
Numérico	Portugal (40)	Itens	.99	.14	.67	1.37	0 (0.0)	0 (0.0)	.89	.40	.22	1.88	3 (8.0)	0 (0.0)
		Sujeitos	1.00	.24	.37	1.76	3 (2.0)	0 (0.0)	.92	1.05	.06	9.09	10 (7.0)	5 (3.5)
	Moçambique (18)	Itens	1.01	.11	.85	1.27	0 (0.0)	0 (0.0)	1.09	.25	.80	1.59	0 (0.0)	0 (0.0)
		Sujeitos	.99	.35	.29	3.14	26 (6.2)	2 (0.5)	1.02	.65	.18	5.08	53 (12.0)	28 (7.0)
	Reino Unido (40)	Itens	.96	.31	.46	1.59	1 (2.5)	0 (0.0)	1.08	1.14	.12	5.45	5 (12.5)	3 (7.5)
		Sujeitos	.98	.26	.50	1.64	1 (3.0)	0 (0.0)	1.03	.82	.11	2.45	7 (20.0)	3 (9.0)
Diagramático	Portugal (40)	Itens	1.01	.13	.81	1.43	0 (0.0)	0 (0.0)	.98	.30	.40	1.89	2 (5.0)	0 (0.0)
		Sujeitos	1.00	.15	.68	1.44	0 (0.0)	0 (0.0)	.95	.42	.23	2.92	12 (8.5)	3 (5.0)
	Moçambique (30)	Itens	1.00	.12	.77	1.37	0 (0.0)	0 (0.0)	1.03	.25	.57	1.80	3 (10.0)	0 (0.0)
		Sujeitos	1.00	.10	.73	1.29	0 (0.0)	0 (0.0)	1.01	.21	.50	2.28	10 (2.4)	2 (0.5)
	Reino Unido (40)	Itens	1.01	.17	.73	1.56	1 (2.5)	0 (0.0)	1.10	.46	.57	3.36	3 (7.5)	1 (2.5)
		Sujeitos	1.00	.14	.73	1.36	0 (0.0)	0 (0.0)	1.03	.35	.44	2.91	8 (8.1)	2 (2.0)
	África do Sul (40)	Itens	1.02	.16	.79	1.42	0 (0.0)	0 (0.0)	1.10	.37	.62	2.39	5 (12.5)	1 (2.5)
		Sujeitos	.99	.14	.60	1.40	0 (0.0)	0 (0.0)	1.00	.32	.37	2.38	17 (5.8)	4 (1.4)
	Austrália (40)	Itens	1.00	.10	.83	1.36	0 (0.0)	0 (0.0)	1.04	.24	.62	1.95	1 (2.5)	0 (0.0)
		Sujeitos	.99	.17	.23	2.04	7 (0.2)	1 (0.0)	.99	.44	.16	4.97	216 (7.6)	81 (2.8)

Relativamente aos índices *outfit*, observa-se também um ajustamento generalizado, com médias *MNSQ* entre .95 e 1.00. Quanto aos sujeitos, também se verificam valores médios *infit* entre .99 e 1.00, e *outfit* entre .97 e 1.00, indicadores de ajustamento satisfatório.

No que diz respeito ao teste Numérico pode-se observar, nos itens, valores médios de *infit* entre .96 e 1.01 e nos sujeitos entre .98 e 1.00. Os índices *outfit* do mesmo teste variam, para os itens, entre .89 e 1.08, e para os sujeitos, entre .92 e 1.03, resultados favoráveis do ponto de vista do ajustamento ao modelo.

Por último, no teste Diagramático, observa-se valores médios de *infit* entre 1.00 e 1.02, para os itens, e entre .99 e 1.00, para os sujeitos. Quanto aos índices *outfit* estes variam, entre .98 e 1.10, para os itens e entre .95 e 1.03 para os sujeitos.

Ainda que os valores médios se enquadrem no modelo, os valores máximos obtidos indicam alguns padrões de respostas inesperados, tanto nos itens como nos sujeitos, sendo que no primeiro caso, os valores efectivos de *infit* não ultrapassam os 2.5% do total de itens e nos de *outfit,* os 12.5%. Quando aos indivíduos, os valores inesperados são, ao nível de *infit* de 6.2% e de *outfit* de 12.0% do total de indivíduos (exceptuando a versão inglesa do teste numérico, cuja percentagem de respostas inesperadas é de 20.0%).

No que concerne o limite mínimo do ajustamento, apenas foram verificados em três versões valores de *infit* dos sujeitos inferiores a .5, contudo representam menos de 1% de valores desajustados.

Torna-se relevante salientar que os índices *infit* são considerados mais robustos que os índices *outfit*, uma vez que são menos sensíveis a *outliers*. Para além disso, é importante interpretar com cautela os valores de ajustamento obtidos pelos sujeitos, pois o número de itens a avaliar os sujeitos é inferior ao número de sujeitos a avaliar os itens.

Presentes na Tabela 2 estão outros indicadores importantes na análise TRI, as estatísticas das pontuações dos itens e dos sujeitos. Nos Modelos logísticos de um parâmetro, como é o caso do Modelo de Rasch, a probabilidade de resposta correcta é obtida através da diferença entre os parâmetros da pessoa e do item (que se expressam na mesma escala intervalar, designada *logit*), onde o ponto 0 corresponde ao nível de dificuldade médio dos itens e, embora tenda para infinito em ambos os extremos, a maioria situa-se entre -5 e +5.

Analisando as pontuações do teste Verbal, pode-se verificar que, de uma forma geral, o nível de competência médio da amostra foi superior ao nível de dificuldade média dos itens e dos sujeitos.

Tabela 2

Análise de Rasch: estatísticas descritivas das pontuações na escala logit

Teste	Versão		Estatísticas das pontuações (*logits*)				
			Min.	Máx.	média	dp	EPmédia
Verbal	Portugal	Itens	-2.85	3.28	.00	1.23	.28
		Sujeitos	-.45	4.09	1.44	.66	.40
	Moçambique	Itens	-1.75	2.34	.00	.91	.14
		Sujeitos	-2.74	2.81	.37	.69	.49
	Reino Unido	Itens	-2.28	2.05	.00	.98	.53
		Sujeitos	-1.29	3.35	1.24	1.16	.39
Numérico	Portugal	Itens	-3.43	5.78	.00	1.79	.46
		Sujeitos	-.86	4.42	1.84	1.08	.71
	Moçambique	Itens	-2.00	1.75	.00	1.10	.22
		Sujeitos	-3.24	2.52	-.50	1.21	.85
	Reino Unido	Itens	-2.93	4.44	.00	1.83	.62
		Sujeitos	-3.85	5.34	.95	1.98	.56
Diagramático	Portugal	Itens	-1.67	2.74	.00	.96	.35
		Sujeitos	-.79	3.71	1.85	.95	.65
	Moçambique	Itens	-.78	1.73	.00	.53	.23
		Sujeitos	-3.34	2.90	-.21	1.17	.68
	Reino Unido	Itens	-1.51	2.17	.00	.81	.30
		Sujeitos	-3.45	2.85	.21	1.26	.50
	África do Sul	Itens	-1.55	1.39	.00	.80	.18
		Sujeitos	-3.22	3.70	.42	1.27	.50
	Austrália	Itens	-1.90	2.17	.00	1.00	.06
		Sujeitos	-3.51	4.15	1.09	1.16	.54

Em relação ao teste Numérico, pode-se verificar que nas versões portuguesa e inglesa o nível médio de atributo é superior ao valor correspondente para os itens. Contudo, na versão moçambicana, o valor médio do nível de competência dos sujeitos é inferior ao valor médio da dificuldade dos itens.

Por último, no teste Diagramático, verifica-se que a pontuação média dos sujeitos é superior à pontuação média dos itens, com excepção da versão moçambicana, onde a pontuação média dos sujeitos é inferior em relação à dificuldade média dos itens. Numa análise mais profunda, pode-se verificar que, de uma forma geral, o nível de competência máximo dos sujeitos é sempre superior ao nível de dificuldade máxima dos itens. Para além disso, nos Mapas de Itens e de Sujeitos (Anexo A), observa-se também que existem poucos itens para diferenciar sujeitos com maior nível de competência.

A análise segundo o Modelo de Rasch fornece ainda informação relativa aos índices de precisão das medidas. Nos Modelos do traço latente, as medidas tradicionais de consistência interna, como o *alfa* de Cronbach ou Kuder-Richardson correspondem, de uma forma geral, ao índice "precisão dos sujeitos" (*person reliability*), referente à probabilidade de uma pessoa obter o mesmo resultado se responder a um conjunto de itens paralelos que

medem o mesmo construto (Bond & Fox, 1998). Este índice desdobra-se em "precisão real" (*real reliability*) e em "precisão do modelo" (*model reliability*), respectivamente os limites inferior e superior do intervalo onde se situa o valor do coeficiente de precisão, sendo que quanto menor o intervalo, menor o ruído presente nos dados (Linacre, 2010). Outro índice importante é o de "precisão do item" (*item reliability*) e diz respeito à replicabilidade da ordenação dos itens, se os mesmos itens fossem apresentados a uma outra amostra da mesma dimensão. Na Tabela 3 estão patentes os coeficientes de precisão dos testes em estudo onde se pode verificar que os coeficientes *alfa* são superiores a .80, com excepção das versões moçambicanas. No que diz respeito à análise de Rasch, observa-se que a diferença entre a "precisão real" e a "precisão do modelo" não ultrapassa .05. É importante referir que, de uma forma geral, o a*lfa* de Cronbach sobrestima a precisão, enquanto a análise de Rasch tende a subestimá-la (Linacre, 2010).

Tabela 3

Análise de Rasch: coeficientes de precisão e erros-padrão

Teste	Versão (nº de itens analisados)	Coeficiente de Precisão				Erro-Padrão	
		Alfa de Cronbach	Sujeitos		Itens	(*logit*)	
			Real	Modelo	Real	m	dp
Verbal	Portugal (56)	.86	.59	.61	.94	.28	.10
	Moçambique (30)	.69	.44	.48	.97	.14	.04
	Inglaterra (60)	.96	.87	.88	.67	.53	.12
Numérico	Portugal (40)	.90	.50	.53	.92	.46	.22
	Moçambique (18)	.56	.49	.54	.95	.22	.11
	Inglaterra (40)	.96	.89	.89	.87	.62	.13
Diagramático	Portugal (40)	.87	.44	.45	.85	.35	.11
	Moçambique (30)	.63	.64	.65	.74	.23	.13
	Inglaterra (40)	.81	.81	.82	.83	.30	.07
	África do Sul (40)	.86	.81	.82	.94	.18	.05
	Austrália (40)	.88	.71	.72	1.00	.06	.01

O método proporciona, além dos índices mencionados, outras informações relativamente à unidimensionalidade, que é um pressuposto de que parte a análise de Rasch e que se torna necessário confirmar se está, de facto, presente. As aptidões medidas através das versões em estudo mostram-se unidimensionais, ainda que os valores da variância explicada pelo modelo estejam aquém do desejável. A unidimensionalidade é adicionalmente comprovada através da análise de contrastes, uma vez que não foram identificados, na variância residual, outros factores que a ponham em causa.

2. Funcionamento Diferencial dos Itens (*DIF*)

Tendo por base os valores de referência referidos no Método ($t > \pm 2.00$, $p < .05$), de seguida analisa-se os três testes, comparando os resultados obtidos nos diversos grupos. Nas tabelas que se seguem, optou-se por identificar apenas os itens que revelam funcionamento diferencial, sendo apresentadas em anexo as tabelas completas (Anexo B).

Comparando as versões portuguesas e moçambicanas dos três testes (Tabela 4), verifica-se que no teste Verbal foram identificados 12 itens que manifestam *DIF*, correspondentes a 40.0% da totalidade de itens no teste, dos quais 26.7% são favoráveis aos

Tabela 4

Comparação entre o grupo de referência (Portugal) e o grupo focal (Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF) nos Testes Verbal, Numérico e Diagramático (versão reduzida)[1][2]

Teste (nº de itens analisados)	Item	Estímulo	Portugal		Moçambique		Contraste *DIF*	*t*	Prob
			Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
Verbal (30)	4	A	-1.28	.36	-.37	.11	-.91	-2.42	.0165
	8	B	-.09	.24	.68	.10	-.77	-2.92	.0038
	10	C	.40	.21	1.21	.11	-.81	-3.35	.0009
	11	C	-.52	.28	.16	.10	-.68	-2.28	.0235
	17	E	-.52	.28	.14	.17	-.66	-2.00	.0461
	33	G	.89	.20	.09	.11	.80	3.48	.0006
	34	G	.37	.22	-.47	.12	.84	3.39	.0008
	40	H	-2.27	.51	-.72	.13	-1.55	-2.98	.0033
	41	I	-.08	.26	1.16	.13	-1.24	-4.26	.0000
	45	J	3.64	.28	2.32	.19	1.32	3.91	.0001
	46	J	3.52	.25	1.12	.15	2.40	8.26	.0000
	56	K	1.25	.30	.12	.22	1.13	2.99	.0034
Numérico (18)	13	e	-3.26	1.01	.50	.16	-3.76	-3.67	.0004
	21	b	-.10	.32	-.81	.13	.71	2.07	.0398
	27	d	1.07	.33	-.16	.21	1.23	3.17	.0019
Diagramático (30)	5	N/A	.08	.25	-.78	.12	.86	3.11	.0021
	8	N/A	-1.71	.46	-.72	.12	-.99	-2.07	.0400
	17	N/A	1.06	.21	-.06	.14	1.13	4.41	.0000
	18	N/A	-.36	.28	-1.00	.14	.64	2.03	.0431
	20	N/A	.84	.21	.33	.14	.52	2.01	.0455
	21	N/A	1.59	.26	.52	.19	1.07	3.35	.0010
	22	N/A	-.84	.34	.36	.17	-1.20	-3.15	.0019
	30	N/A	1.69	.30	.61	.34	1.07	2.38	.0191

Nota: N/A – Não se aplica

[1] As versões completas encontram-se em anexo (Anexo B). A coluna "Estímulo" diz respeito ao texto a que o item reporta, no caso do teste Verbal, ou à tabela/gráfico, no caso do teste Numérico.

[2] Teste Verbal: Portugal (n=139), Moçambique (n=422); teste Numérico: Portugal (n=139), Moçambique (n=418); teste Diagramático: Portugal (n=141), Moçambique (n=419).

respondentes portugueses, isto é, para que se acerte em cada um destes itens é necessário um menor nível de competência na amostra portuguesa, comparativamente à amostra moçambicana. Torna-se relevante salientar que entre estes, há dois casos cujos itens reportam ao mesmo estímulo (itens 10 e 11, itens 33 e 34).

Relativamente ao teste Numérico, verifica-se a presença de *DIF* em apenas 3 itens (16.7% da total de itens), dos quais 2 são mais fáceis para a amostra moçambicana. Entre os itens identificados, não existe nenhum estímulo em comum.

Por fim, comparando as versões portuguesa e moçambicana no teste Diagramático, identificou-se a presença de *DIF* em 8 itens, que representam 26.7% da total de itens no teste, dos quais 20.0% são mais fáceis para à amostra moçambicana.

Na Tabela 5 são identificados os itens que manifestaram *DIF* entre a amostra portuguesa e inglesa nos testes Verbal, Numérico e Diagramático.

Comparando o teste Verbal na amostra portuguesa e inglesa, pode-se observar que 5 itens apresentam *DIF*, correspondentes a 19.2% dos itens em comum entre as duas versões, sendo que 3 itens são mais fáceis para a amostra portuguesa. No conjunto destes itens, há dois estímulos em comum (itens 17 e 19, itens 33 e 35).

Relativamente ao teste Numérico, verifica-se que 13 itens manifestam *DIF*, representando 32.5% do total de itens do teste, dos quais 8 são mais fáceis para a amostra portuguesa. Neste caso, importa salientar a presença de 4 itens que reportam ao mesmo estímulo e, embora não se revelem favoráveis a um grupo específico, manifestam diferença no funcionamento do item nas duas amostras (itens 6, 14, 18 e 25). Um outro conjunto de 3 itens (15, 22 e 34) referem-se a outro estímulo, embora neste caso são todos favoráveis à amostra portuguesa.

Por último, no teste Diagramático, foi identificado *DIF* em 6 itens (15% do total do teste), embora não tenha sido identificada uma amostra cuja percentagem de itens fosse mais favorável.

Após a comparação e a identificação de *DIF* nos três testes em estudo, entre a amostra inglesa e portuguesa, e entre a amostra portuguesa e moçambicana, revelou-se interessante o estudo do teste Diagramático com outras amostras, visto que os seus estímulos são isentos de conteúdos verbais ou numéricos. Assim, para os efeitos do presente estudo, apresenta-se de seguida a comparação separada entre a versão inglesa e a versão portuguesa e as restantes versões.

Tabela 5

Comparação entre o grupo de referência (Portugal) e o grupo focal (Reino Unido) para a detecção de Funcionamento Diferencial dos Itens (DIF) nos Testes Verbal, Numérico e Diagramático (versão reduzida)[3]

Teste (nº de itens analisados)	Item	Estímulo	Portugal		Reino Unido		Contraste *DIF*	*t*	Prob
			Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
Verbal (26)	17	D	-.78	.29	.54	.48	-1.33	-2.37	.0211
	19	D	.03	.22	1.44	.46	-1.41	-2.77	.0077
	26	F	-1.11	.33	.21	.49	-1.33	-2.23	.0288
	33	G	.60	.20	-1.12	.66	1.73	2.50	.0169
	35	G	3.21	.24	2.05	.45	1.15	2.27	.0267
Numérico (40)	6	f	.44	.22	-1.98	.55	2.42	4.07	.0001
	8	h	-.05	.26	-1.51	.59	1.46	2.28	.0264
	13	e	-3.31	.81	-1.20	.62	-2.11	-2.06	.0419
	14	f	-1.42	.40	-.13	.47	-1.29	-2.08	.0402
	15	g	-.48	.35	.92	.46	-1.41	-2.43	.0175
	18	f	-.74	.32	.44	.46	-1.17	-2.10	.0394
	22	g	-1.14	.43	1.42	.46	-2.56	-4.07	.0001
	23	c	-.38	.32	.79	.46	-1.17	-2.09	.0401
	25	f	1.32	.24	-1.00	.52	2.32	4.02	.0002
	27	d	.58	.29	-1.18	.58	1.76	2.72	.0085
	34	g	.78	.67	3.30	.72	-2.51	-2.55	.0168
	36	h	1.52	.72	4.18	.78	-2.66	-2.50	.0196
	38	b	2.89	.59	-1.02	.56	4.00	4.89	.0000
Diagramático (40)	2	N/A	-.82	.31	.47	.24	-1.29	-3.26	.0013
	17	N/A	1.03	.21	.25	.24	.78	2.41	.0168
	22	N/A	-.86	.33	.54	.27	-1.40	-3.27	.0013
	27	N/A	-3.14	.69	-.55	.29	-2.59	-3.45	.0007
	38	N/A	3.27	.98	.56	.45	2.71	2.52	.0306
	39	N/A	2.72	.61	.45	.45	2.27	2.99	.0053

Nota: N/A – Não se aplica

Na comparação do funcionamento dos itens do teste Diagramático entre a amostra portuguesa e as amostras australiana e sul-africana (Tabela 6), verifica-se que em ambos os casos 5 itens manifestam *DIF* (12.5% do total de itens), sendo que na comparação entre a versão portuguesa e a versão australiana, há 4 itens mais fáceis em Portugal. Na Tabela 7 apresentam-se os itens que manifestam *DIF* entre a versão inglesa e as versões australiana, sul-africana e moçambicana. Relativamente à primeira, foram identificados 11 itens com *DIF* (27.5% do total do teste), dos quais 6 favoráveis à versão inglesa; na segunda identificaram-se apenas 2 itens (5.0% do teste); e na terceira foram identificados 4 itens (13.3% do teste), embora não sejam favoráveis a uma das culturas.

[3] Teste Verbal: Portugal (n=139), Reino Unido (n=29); teste Numérico: Portugal (n=139), Reino Unido (n=32); teste Diagramático: Portugal (n=141), Reino Unido (n=98).

Tabela 6

Comparação entre o grupo de referência (Portugal) e os grupos focais (Austrália e África do Sul) para a detecção de Funcionamento Diferencial dos Itens (DIF) no Teste Diagramático (versão reduzida)[4]

Item	Portugal		Austrália		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
6	-.31	.26	-.95	.06	.64	2.40	.0176
16	-.42	.28	.37	.05	-.79	-2.77	.0064
24	-.64	.33	.23	.05	-.87	-2.60	.0104
27	-3.14	.69	-1.26	.07	-1.88	-2.71	.0079
34	-.46	.51	.88	.07	-1.34	-2.60	.0130

Item	Portugal		África do Sul		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
2	-.82	.31	-.05	.14	-.77	-2.23	.0265
22	-.86	.33	.28	.15	-1.14	-3.14	.0020
27	-3.14	.69	-1.16	.19	-1.98	-2.76	.0066
38	3.27	.98	.29	.28	2.98	2.93	.0189
39	2.72	.61	1.28	.32	1.44	2.09	.0454

Tabela 7

Comparação entre o grupo de referência (Reino Unido) e os grupos focais (Austrália, África do Sul e Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF) no Teste Diagramático (versão reduzida)[4]

Item	Reino Unido		Austrália		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
2	.47	.24	-.66	.05	1.13	4.60	.0000
4	-1.47	.27	-.91	.06	-.56	-2.01	.0467
7	.20	.23	-.61	.05	.81	3.41	.0008
10	-.49	.24	-1.06	.06	.57	2.31	.0222
17	.25	.24	.92	.05	-.66	-2.69	.0081
22	.54	.27	-.31	.05	.85	3.08	.0027
27	-.55	.29	-1.26	.07	.71	2.37	.0196
30	.85	.32	1.76	.06	-.91	-2.79	.0067
34	-.10	.35	.88	.07	-.98	-2.72	.0088
36	.42	.40	1.42	.07	-.99	-2.42	.0196
39	.45	.45	1.63	.09	-1.18	-2.56	.0142

Item	Reino Unido		África do Sul		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
10	-.49	.24	-1.08	.16	.60	2.07	.0392
24	-.70	.30	.06	.17	-.76	-2.21	.0290

Item	Reino Unido		Moçambique		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
4	-1.52	.28	-.77	.12	-.75	-2.47	.0144
5	.02	.24	-.78	.12	.80	2.90	.0042
14	-1.29	.27	-.61	.13	-.68	-2.27	.0244
21	1.63	.35	.52	.19	1.12	2.81	.0057

[4] Teste Diagramático: Portugal (n=141), Moçambique (n=419); Reino Unido (n=98); Austrália (n=2817); África do Sul (n=291)

Concluindo a apresentação dos resultados, revela-se interessante salientar algumas particularidades relativas ao funcionamento dos itens, paralelas às diversas culturas. No que diz respeito ao teste Verbal, verifica-se que os itens que reportam ao estímulo "G" manifestam *DIF*, tanto entre a versão inglesa e portuguesa, como entre a portuguesa e a moçambicana. Por outro lado, há a salientar o funcionamento do item 22 do teste Diagramático, que se revela mais fácil para a amostra portuguesa em comparação com a amostra moçambicana, inglesa e sul-africana, assim como o item 27, desta feita comparando com a amostra inglesa, australiana e sul-africana.

Outro aspecto que merece destaque é referente aos dois itens do teste Diagramático que são mais fáceis para Moçambique por comparação com o Reino Unido, e que foram igualmente identificados quando se comparou esta versão com a versão portuguesa. Assim, para responder correctamente aos itens 5 e 21, é necessário um menor nível de competência em Moçambique, do que em Portugal ou no Reino Unido.

Discussão

Ajustamento ao Modelo de Rasch

Previamente à discussão da hipótese proposta, é importante atender aos resultados obtidos nas análises separadas de cada versão em estudo. Como mencionado anteriormente, os dados das diferentes distribuições de resultados ajustam-se muito satisfatoriamente ao Modelo de Rasch, sendo esse o ponto de partida para as análises e interpretações subsequentes.

Do ponto de vista da consistência interna (estudo de precisão) verifica-se que, de uma forma global, os coeficientes *alfa* de Cronbach obtidos podem ser considerados bons. Porém, os coeficientes emergentes da análise TRI não são muito elevados e antecipam desde logo algumas limitações da qualidade dos dados. Embora não haja na literatura um valor de referência que se considere como mínimo desejável, o facto de não serem muito elevados constitui advertência para a interpretação cautelosa dos restantes resultados.

Funcionamento Diferencial dos Itens (*DIF*)

O presente estudo procurou analisar a equivalência entre cinco versões de uma bateria de testes de aptidões, através do estudo do *DIF*, sendo que a eventual existência deste tipo de enviesamento questiona a correspondência intercultural do teste utilizado. Como tal, e para discutir os resultados obtidos à luz da hipótese proposta, considera-se como versão original a

versão inglesa, quando comparada com a versão portuguesa, australiana e sul-africana, e a versão portuguesa quando se compara com a versão moçambicana.

Na comparação do teste Verbal, entre a versão portuguesa e a inglesa foi identificado *DIF* em 5 itens, embora esse enviesamento seja equilibrado em itens favoráveis a uma e a outra amostra. É de salientar que estas versões só partilham 26 itens, ou seja, cerca de metade da versão original (inglesa). Assim, na adaptação portuguesa, a análise realizada pelos seus construtores, através do Método dos Juízes, terá levado a eliminar itens difíceis de traduzir ou pouco adequados à língua ou cultura portuguesa.

Por sua vez, entre a versão portuguesa e a moçambicana, foi identificado um número superior de itens favoráveis a Portugal. Uma possível justificação deste resultado pode estar relacionada com a aplicação do Método dos Juízes, uma vez que, ao contrário do exemplo anterior, a construção da versão moçambicana foi realizada em Portugal, logo não foi efectuada no país a que se destinava o teste. Este caso sugere que o Método dos Juízes deve ser aplicado recorrendo a especialistas inseridos na cultura de cada país, de forma a minimizar o eventual impacto negativo da utilização de testes pouco adaptados ao contexto.

Na comparação entre a versão portuguesa e a inglesa no teste Numérico, foram identificados vários itens com funcionamento diferente, alguns deles reportando ao mesmo estímulo, o que sugere que a tradução do próprio gráfico ou tabela poderá ter condicionado o seu funcionamento adequado. Pelo contrário, na comparação entre a versão portuguesa e a moçambicana, poucos itens manifestaram *DIF*. Uma vez que a versão moçambicana é composta por 18 itens, resultantes da escolha entre os 40 itens da versão portuguesa, pode-se concluir que a selecção destes itens foi mais eficaz.

Embora a análise dos itens e a identificação de *DIF* seja específica do teste em questão, várias investigações internacionais foram conduzidas para estudar a equivalência de testes em diferentes países, tanto ao nível dos testes verbais (Allalouf & Sireci, 1998; Ellis, 1989), como ao nível dos testes numéricos (Fox & Verhagen, 2011; Yildirim, 2006). Dentro dos poucos estudos que têm relacionado o conteúdo do item e o *DIF*, destaca-se o trabalho de Ellis (1989) que concluiu que a maioria dos itens em que se detecta *DIF*, este efeito resulta de erros de tradução. Ainda neste sentido, Hulin (1987) sugere um método para explicar as fontes de *DIF* utilizando os parâmetros da TRI. Segundo este autor, diferenças no parâmetro da dificuldade do item resultariam de erros na sua tradução, enquanto diferenças no parâmetro de discriminação estariam relacionados com a relevância cultural do item. Deste modo, de acordo com a teoria de Hulin, seria necessária também uma análise do parâmetro da discriminação para concluir efectivamente que itens seriam enviesados. Para testar esta

hipótese, teria de ser aplicado um Modelo TRI de dois parâmetros para estudar a dificuldade e a discriminação de cada item.

Os testes verbais e os testes numéricos têm conteúdos que são sujeitos a traduções, podendo inclusive ser, como já referido, a própria causa do *DIF*. O presente estudo beneficia do facto de poder apresentar também dados relativos a um instrumento isento de conteúdos semânticos. Porém, o teste Diagramático também apresenta *DIF* nas várias comparações efectuadas, não podendo ser justificado através da tradução ineficaz dos itens, pois ela não existe, nem pela selecção errada por parte dos Juízes. Assim, as diferenças encontradas poderão estar relacionadas com diferenças entre os vários países, no próprio raciocínio lógico. De entre as várias comparações, destaca-se as diferenças encontradas entre Portugal e Moçambique, onde há um conjunto superior de itens favoráveis à versão moçambicana, sendo este dado curioso uma vez que o nível de competência médio desta população neste teste é bastante inferior ao nível de competência médio da população portuguesa no mesmo teste (ver Tabela 2). Nesta versão, foram estudados apenas 30 itens, alguns favoráveis à amostra moçambicana, mas não se sabe o comportamento dos 10 itens que não foram escolhidos (dos 40 itens totais do teste Diagramático) e que poderiam beneficiar, em parte, a população portuguesa e de alguma forma equilibrar o número de itens com *DIF* no teste. Um outro aspecto que poderá condicionar este resultado, e que será aprofundado posteriormente, é o elevado número de respostas omissas nos testes moçambicanos.

Outros dados que corroboram a hipótese de que o *DIF* identificado resulta das diferenças no raciocínio lógico entre as amostras, decorrem das comparações do teste Diagramático com a amostra australiana e sul-africana. Embora tenham sido identificados itens que possuem um comportamento diferente transversalmente entre versões, a maioria dos itens que manifestam *DIF* são sempre diferentes nas várias comparações, não sendo possível reconhecer um padrão no seu funcionamento.

No presente estudo foram identificados itens que manifestam um funcionamento diferencial (*DIF*), logo a hipótese proposta foi refutada, ou seja, a equivalência entre as versões estudadas é posta em causa devido a existência de itens enviesados. Porém, a identificação de *DIF* não permite concluir que o teste na sua globalidade é culturalmente enviesado, uma vez que não foi realizado um estudo do Funcionamento Diferencial do Teste (*DTF*). Ao adoptar o método proposto por Raju, van der Linden e Fleer (1995), que permite que os efeitos do *DIF* sejam compensatórios, isto é, o tamanho do *DIF* de um item pode ser anulado pelo tamanho do *DIF* de outro item, pode-se averiguar a possibilidade de existir itens enviesados mas o teste ser equivalente.

Limitações

Na presente investigação, foram identificadas algumas limitações, sendo as principais relacionadas com a caracterização da amostra, nomeadamente ao nível da omissão de dados demográficos. A falta de conhecimento de variáveis de controlo, como as habilitações literárias ou as áreas profissionais, impede a interpretação do *DIF* manifestado pelos itens como resultado de sensibilidades culturais, pois pode decorrer em vez disso de características demográficas das amostras em causa. Um outro aspecto que deve ser também mencionado, embora relacionado com o tópico anterior, diz respeito ao real controlo da nacionalidade dos participantes, na medida em que poderão existir indivíduos que respondem a uma versão do teste e não são naturais do país em questão.

Outra limitação deste estudo consiste na dimensão da amostra e no facto de não haver uma distribuição equitativa pelo nível de competência. Entre as várias amostras, é visível a sobre-representação da Austrália, bem como a sub-representação da amostra do Reino Unido, o que condiciona os resultados obtidos em diversas análises. Embora não haja um valor mínimo para a detecção de *DIF*, quanto menor a amostra, menos sensível é o *DIF* (Clauser & Mazor, 1998; Linacre, 2010). Como aconselham Hambleton e Jones (1994), é importante interpretar com cautela resultados de amostras de reduzida dimensão, como é o caso das comparações entre as versões portuguesas e inglesas. Aliás, Linacre (2010) afirma que, quando um grupo é inferior a 30 elementos, há muita influência de comportamentos idiossincráticos que podem afectar o *DIF*, sendo esta a razão que leva alguns autores a sugerirem que as amostras sejam constituídas por grupos entre os 200 e 250 elementos (Clauser & Mazor, 1998). Outro aspecto que merece destaque é a falta de uma distribuição regular ao longo do nível de competência, uma vez que alguns índices da TRI (como é o caso da "precisão dos sujeitos") são influenciados por esta distribuição. Esta contingência é também a razão pela qual o presente estudo não apresenta os resultados Mantel-Haenszel, os mais conhecidos na literatura do *DIF* (Clauser & Mazor, 1998, Emenogu, Falenchuk. & Childs, 2010; Hambleton & Jones, 1994; Hambleton & Rogers, 1989; Holland & Thayer, 1998; Linacre, 2010). Embora haja consenso de que os itens identificados por este método são também identificados pelo método da TRI (Hambleton & Jones, 1994; Hambleton & Rogers, 1989), nomeadamente quando os dados se ajustam ao modelo (Linacre, 2010), o Método Mantel-Haenszel exige que haja poucas respostas omissas e muitos sujeitos nos vários níveis de competência.

Como já mencionado, o presente estudo conta com muitas respostas omissas, principalmente na amostra moçambicana. Finch (2008) refere que em estudos de avaliação

cognitiva, as respostas omissas podem condicionar a estimação correcta dos parâmetros utilizados. Sobre este assunto, Mislevy e Wu (1988) referem que a falta de respostas não condiciona a qualidade da estimação dos parâmetros, e afirmam inclusive que se não forem incluídos os dados omissos, torna-se possível que haja enviesamento nos parâmetros dos itens e dos sujeitos. Emenogu e colaboradores (2010) acrescentam que quando há poucas respostas omissas, o tratamento como omissas não vai influenciar o resultado, contudo à medida que este número aumenta, pode influenciar a estimação e, em última instância, ser a própria causa do *DIF*.

Outros aspectos a ter em consideração estão relacionados com a ordenação dos próprios itens. Através dos Mapas de Itens e de Sujeitos (Anexo A), observa-se que a ordenação dos itens não respeita os níveis de dificuldade dos mesmos, ou seja, há itens mais fáceis após itens mais difíceis. Este aspecto não seria uma limitação se todas as versões tivessem a mesma ordenação, estando todos os sujeitos em igualdade de circunstâncias. Porém, as versões de cada país não tem a mesma ordenação (com excepção do teste diagramático em Portugal, Reino Unido, Austrália e África do Sul), pelo que podem existir sujeitos que seriam capazes de responder a um dado item por estar dentro do seu nível de competência, mas na versão que responderam o item situa-se numa posição em que o sujeito não teve oportunidade de o responder.

A última limitação a enumerar relaciona-se com o controlo dos outros tipos de enviesamentos definidos por van de Vijver e Leung (1997): Enviesamento do Construto e Enviesamento do Método. Os testes utilizados neste estudo, foram construídos de acordo com a segunda abordagem definida por Anastasi e Urbina (1997) ou "*Pseudoetic*" (Davidson et al., 1976), uma vez que o teste foi construído no contexto inglês, e aplicado posteriormente às outras culturas, o que não garante automaticamente a Equivalência de Construto (van de Vijver & Leung, 1997). Adicionalmente, não há garantias de que não houve factores de ruído, como por exemplo nas condições de aplicação do instrumento nas diversas amostras, os quais condicionam o Enviesamento do Método.

Sugestões para futuras investigações

As sugestões apresentadas visam colmatar as limitações identificadas previamente. Em primeiro lugar, sugere-se que em futuras investigações seja assegurado o controlo das variáveis demográficas, tendo em vista a obtenção de conclusões mais fidedignas e com maior impacto nos construtores e utilizadores de testes. Paralelamente, amostras de maior dimensão

permitem a estimação mais precisa dos parâmetros e dos índices da TRI, consolidando assim as conclusões obtidas.

Uma outra sugestão refere-se à reordenação dos itens segundo os parâmetros de dificuldade, mas principalmente sobre a transversalidade da ordenação dos itens nas versões em estudo, de forma a colocar todos os respondentes em igualdade de circunstâncias.

Relativamente à tipologia de enviesamento do item proposta por Mellenberg (1982, cit. por van de Vijver & Poortinga, 2005), os dados obtidos apenas permitem estimar o *DIF* para itens uniformes. Em futuras investigações, seria interessante estudar o *DIF* em itens não uniformes, cuja diferença entre os grupos varia ao longo do nível de competência, sendo neste caso novamente necessária uma amostra equilibradamente distribuída pela dimensão medida (Linacre, 2010).

No que diz respeito ao número de respostas omissas verificadas na presente amostra (principalmente na amostra moçambicana), Emenogu e colaboradores (2010) sugerem a eliminação de respostas omissas, tendo como objectivo o aumento da proporção de itens respondidos.

Como mencionado no enquadramento teórico, vários investigadores enumeram as vantagens da TRI relativamente à TCT (Embretson & Reise, 2000; Fan, 1998). Uma das vantagens da TRI relaciona-se com a utilização desta metodologia na organização de bancos de itens (Andriola, 1998; Embretson & Reise, 2000; Pasquali & Primi, 2003), que permitem o desenvolvimento de testes adaptativos (*Computerized Adaptative Testing*) através de algoritmos de respostas, reduzindo o tempo na avaliação e aumentando a precisão da estimação do nível de competência dos participantes. Para além disso, por ser aplicado por meio de computador, permite também um maior controlo sobre o Enviesamento do Método.

A maioria dos estudos relativos ao enviesamento centra-se no *DIF*, que analisa o funcionamento de cada item individualmente, descurando o funcionamento global do teste, assim como do próprio indivíduo. Porém, na literatura da TRI, há dois conceitos que poderão ser estudados em futuras investigações: o Funcionamento Diferencial do Sujeito (*Differential Person Functioning – DPF*), que coincide, em alguns aspectos, com o estudo dos itens não uniformes (Linacre, 2010), e o Funcionamento Diferencial do Teste (*Differential Test Functioning – DFT*), que analisa de uma forma holística o conjunto dos itens que constituem o teste e permite determinar o efeito que a adição ou a remoção de itens com DIF pode ter na análise global do seu funcionamento (Drasgow & Probst, 2005).

Implicações práticas

O presente projecto de investigação aplicada difere da investigação fundamental, por estudar aprofundadamente uma bateria de testes de aptidões específica, com elevada utilização em contexto de avaliação psicológica nas organizações. Devido a este facto, o presente estudo tem algumas implicações práticas para a utilização dos testes que constituem esta bateria, principalmente porque foi identificado *DIF* nalguns itens que reportam a determinados textos, tabelas ou gráficos. Neste sentido, seria vantajoso conduzir uma reavaliação dos mesmos nas várias versões, através de uma análise aprofundada das possíveis causas do enviesamento, ainda que a falta de dados demográficos não permita discernir em definitivo se existe efectivamente enviesamento cultural. No caso de se verificar *DIF*, seria clara a necessidade de proceder à modificação dos itens.

Nesta reavaliação dos itens é importante recorrer a uma combinação de Métodos Estatísticos e de Juízes, visto que esta investigação também alerta os construtores dos testes para que uma equivalência eficaz deverá passar por essa mesma conjugação.

Importa também reflectir acerca da implicação do presente estudo para as práticas de recursos humanos, que recorrem frequentemente a instrumentos de avaliação psicológica. Como mencionado nas directrizes da *ITC* (2010), é da responsabilidade dos utilizadores de testes a interpretação cautelosa dos resultados obtidos pelos participantes e da integração de informações dos respectivos contextos socioculturais. No entanto, dado ao aumento de países com populações culturalmente heterogéneas (Berry, Poortinga, Breugelmans, Chasiotis, & Sam, 2011) é também da responsabilidade dos utilizadores assegurar que não existe enviesamento dos itens e que o instrumento que utilizam é metodologicamente equivalente à versão original, garantindo que não contribui para a discriminação injusta de pessoas pertencentes a variados grupos sociais. Neste sentido, poderia adoptar-se o exemplo da África do Sul, onde as técnicas de avaliação psicológica são obrigatoriamente certificadas pelo chamado *Employment Equity Act*, para que os procedimentos de selecção não discriminem injustamente membros de determinados grupos (Theron, 2007).

Embora seja uma temática pouco aprofundada no panorama português, o presente estudo demonstra também a valência da metodologia TRI nos estudos interculturais, pois permite a obtenção de resultados mais robustos e consequentemente analisar de forma aprofundada os próprios itens, para posteriormente trabalhá-los no sentido de causarem um menor enviesamento cultural dos resultados de medida. Comprova também as vantagens desta metodologia, por comparação com os métodos clássicos de estudos dos itens, e serve igualmente como estímulo para uma maior utilização na investigação.

Referências Bibliográficas

Allalouf, A., & Sireci, S. (Abril, 1998). *Detecting sources of DIF in translated verbal items*, Paper apresentado na Reunião Anual da American Educational Research Association, California.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing.* New Jersey: Prentice.

Andriola, W. B. (1998). Utilização da teoria da resposta ao item (TRI) para a organização de um banco de itens destinados a avaliação do raciocínio verbal. *Psicologia, Reflexão e Crítica, 11*, 1-13.

Bartram, D. (2004). Assessment in Organisations. *Applied Psychology: An International Review, 53*(2), 237-259. doi:10.1111/j.1464-0597.2004.00170.x.

Berry, J. W., Poortinga, Y. H., Breugelmans, S. M., Chasiotis, A., & Sam, D. L. (2011). *Cross-cultural psychology: Research and applications*. New York: Cambridge University Press.

Bond, T., & Fox, C. (2007). *Applying the rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Byrne, B., Oakland, T., Leong, F., van de Vijver, F., Hambleton, R., Cheung, F., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology*, *3*(2), 94-105. doi:10.1037/a0014516.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures for identify differential functioning test items. *Educational Measurement*: *Issues and Practice, 17,* 31-44.

Davidson, A., Jaccard, J., Triandis, H., Morales, M., & Diaz-Guerrero, R. (1976). Cross-cultural model testing: toward a solution of the etic-emic dilemma. *International Journal of Psychology, 11*(1), 1-13.

Dickes, P., Tournois, J., Flieller, A & Kop, J-L. (1994). *La psychométrie: théories et méthodes de la mesure en psychologie*. Paris: Presses Universitaires de France.

Drasgow, F., & Probst, T. H. (2005). The psychometrics of adaptation: Evaluating measurement equivalence across languages. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 265-296). Mahwah, NJ: Lawrence Erlbaum Associates.

Duarte, M. E., & Rossier, J. (2008). Testing and assessment in na international context: Cross-and multi-cultural issues. In J. Athanasou & R. Van Esbroeck (Eds.), *International Handbook of Career Guidance* (pp. 489-510). New York: Springer.

Ellis, B. B. (1989). Differential item functioning: implications for test translation. *Journal of Applied Psychology, 74* (6), 912-921.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Emenogu, B. C., Falenchuk, O., & Childs, R. A. (2010). The effects of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research, 56*(4), 459-469.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 1-17.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245.

Fox, J. P., & Verhagen, J. (2011). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet, *Cross-cultural analysis: Methods and applications* (pp. 461-482). New York: Taylor & Francis Group.

Gomes, J., Cunha, M., Rego, A., Cunha, R., Cabral-Cardoso, C., & Marques, C. (2008). *Manual de Gestão de Pessoas e do Capital Humano*. Lisboa: Edições Sílabo.

Gregoire, J., & Hambleton, R. (2009). Advances in test adaptation research: A special issue. *International Journal of Testing, 9*(2), 75-77. doi:10.1080/15305050902880678.

Hambleton, R. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 253-262.

Hambleton, R. K., & Jones, R. W. (1994) Comparison of empirical and judgmental methods for detecting differential item functioning. *Education Research Quarterly, 18*(1), 21-36.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, 11(3), 147-157. doi:10.1027/1015-5759.11.3.147

Hambleton, R. K., & Patsula, L. (1999) Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1-30. Retirado de http://data.memberclicks.com/site/atpu/volume%201%20issue%201Increasing%20validity.pdf

Hambleton, R. K., & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2*(4), 313-334.

Holland, P.W., & Thayer, D.T. (1988) Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C. H. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18*(2), 115-142.

International Test Commission. (2010). International Test Commission Guidelines for Translating and Adapting Tests. Retirado de http://www.intestcom.org

Linacre, J.M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing*. Princeton, NJ: Educational Testing Service.

Moreira, J. (2004). *Questionários: Teoria e prática*. Coimbra: Almedina.

Muñiz, J. (2010). Las teorias de los tests: Teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 3*(1), 57-66.

Pasquali, L., & Primi, R. (2003). Fundamentos da teoria da resposta ao item (TRI). *Avaliação Psicológica, 22*, 99-110.

Prieto, G., & Almeida, L. (1997). Equivalência de pontuações nos testes: Uma solução psicométrica para o dilema émico-ético na avaliação psicológica. *Psicologia: Teoria, Investigação e Prática, 2*, 19-30.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-Baser internal measures of differential functioning of items and tests. *Applied Psychological Measurement*; *19*; 353-368. doi: 10.1177/014662169501900405

Rust, J., & Golombok, S. (1999). *Modern psychometrics: The science of psychological assessment*. London: Routledge.

Ryan, A., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: nation and culture as explanations for variability in practice. *Personnel Psychology, 52*(2), 359-391.

Scheuneman, J. D., & Bleinstein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2*(3), 255-275.

SHL. (1991a). *Bateria de testes CRTB (Manual e guia do utilizador)*. Lisboa: Autor.

SHL. (1991b). *Critical Reasoning Test Battery (Manual and user's guide)*. Surrey, United Kingdom: Author.

Sireci, S., Patsula, L., & Hambleton, R. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Sousa, A. (1999). A selecção e os testes psicológicos: Situação actual e perspectivas futuras no âmbito da consultoria organizacional. In A. P. Soares, S. Araújo & S. Caires (Eds.) Avaliação psicológica: Formas e contextos (pp. 123-134). Braga: APPORT

Spearman, C. (1904) General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

Theron, C. (2007). Confessions, scapegoats and flying pigs: psychometric testing and the law. *SA Journal of Industrial Psychology, 33*(1), 102-117.

Urbina, S. (2004*). Essentials of Psychological Testing*. New Jersey: John Wiley & Sons.

van de Vijver, F. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

van de Vijver, F. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37. doi:10.1027/1015-5759.13.1.29

van de Vijver, F. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*(2), 119-135. doi:10.1016/j.erap.2003.12.004

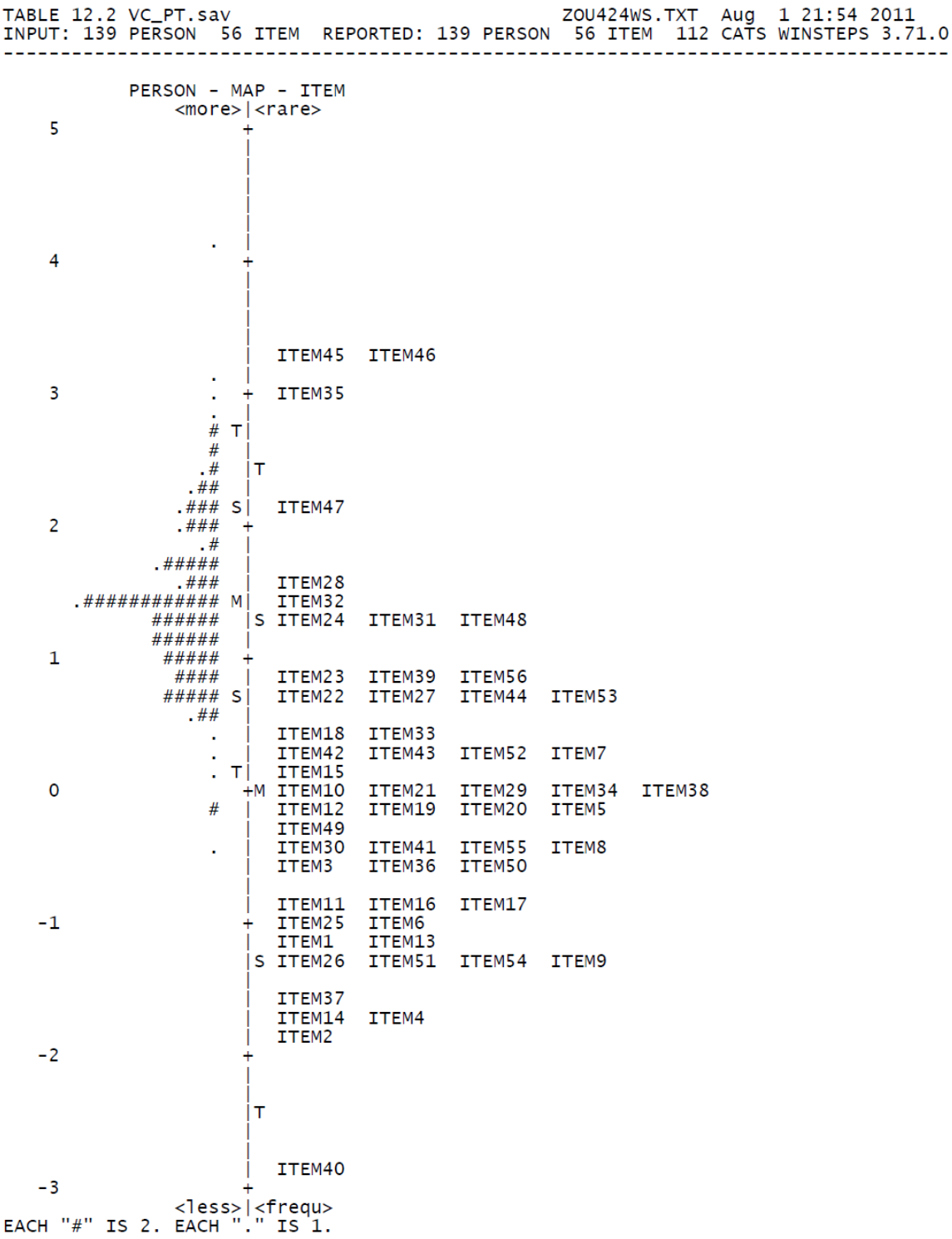
Yildirim, H. H. (2006). *The differential item functioning (DIF) analysis of mathematics items in the international assessment programs* (Dissertação de Doutoramento não publicada). Abant Izzet Baysal University, Turquia.

Anexos

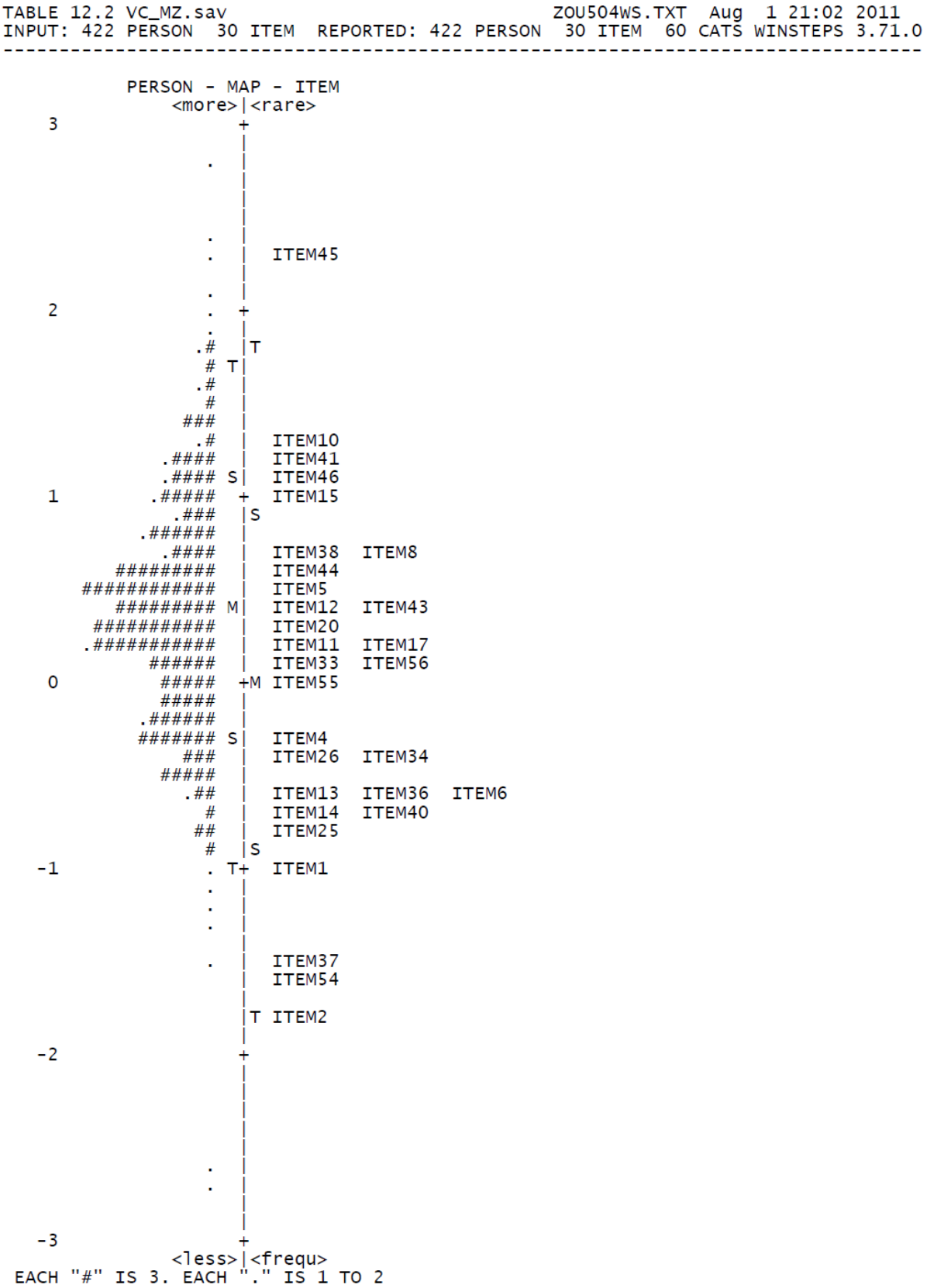
Anexo A

Mapa de Itens e de Sujeitos

Teste Verbal, versão portuguesa: Mapa de Itens e de Sujeitos



Teste Verbal, versão moçambicana: Mapa de Itens e de Sujeitos



Teste Verbal, versão inglesa: Mapa de Itens e de Sujeitos

```
TABLE 12.2 VC_UK.sav                              ZOU872WS.TXT  Aug  1 21:58 2011
INPUT: 29 PERSON  26 ITEM  REPORTED: 29 PERSON  26 ITEM  52 CATS WINSTEPS 3.71.0
-------------------------------------------------------------------------------

     PERSON - MAP - ITEM
         <more>|<rare>
    4        X  +
               T|
                |
                |
                |
          XXX   |
                |
                |
                |
                |
    3           +
                |
                |
          XXX S|
            X   |
                |
                |
           XX   |
                |
                |
    2           +  ITEM35
          XXX   |
                |
                |
           XX   |T
              M|
           XX   |  ITEM19
                |
                |
                |
    1           +
         XXXX   |
                |S ITEM38
           XX   |
                |  ITEM17
                |  ITEM23  ITEM52
           XX   |
              S|  ITEM11  ITEM12  ITEM41  ITEM42
            X   |  ITEM25  ITEM26  ITEM27
                |  ITEM22  ITEM5   ITEM6
    0           +M
                |
                |  ITEM34
                |  ITEM50
                |
            X   |  ITEM18  ITEM37  ITEM8
            X   |  ITEM51
                |
                |S
            X   |  ITEM43
   -1         T+
                |  ITEM33
                |
                |
                |
                |  ITEM49
                |T
                |  ITEM9
                |
                |
   -2           +
         <less>|<frequ>
```

Teste Numérico, versão portuguesa: Mapa de Itens e de Sujeitos

```
TABLE 12.2 NC_PT.sav                                ZOU208WS.TXT  Aug  1 20:54 2011
INPUT: 139 PERSON  40 ITEM  REPORTED: 139 PERSON  39 ITEM  78 CATS WINSTEPS 3.71.0
-------------------------------------------------------------------------------

           PERSON - MAP - ITEM
               <more>|<rare>
    6    XXXXXXXXXX  +
                     |  ITEM37
                     |
                     |
                     |
                     |
    5                +
                     |
                     |
                  X  |
                 XX  |
                  X  |
    4              X T+  ITEM38
                     |
                XXX  |
               XXXX  |T
                 XX  |
                     |
    3              X  +
                XXXX S|
           XXXXXXXX  |
           XXXXXXXX  |
         XXXXXXXXXX  |  ITEM26  ITEM33
       XXXXXXXXXXXX  |
    2          XXXXX  +  ITEM35
          XXXXXXXXX M|S
            XXXXXXX  |
           XXXXXX    |  ITEM28  ITEM36
           XXXXXX    |  ITEM25
          XXXXXXXX   |
    1           XXX  +  ITEM31
              XXXXX S|  ITEM7
                  X  |  ITEM11  ITEM27  ITEM29  ITEM34
              XXXXX  |  ITEM6
               XXXX  |
                 XX  |  ITEM20  ITEM5
    0             X  +M ITEM8
                  XX  |  ITEM19  ITEM24  ITEM30  ITEM39
                     T|  ITEM21  ITEM23
                      |  ITEM15  ITEM32
                  X   |
                 XXX  |  ITEM18  ITEM3
   -1                 +
                      |  ITEM22
                      |  ITEM16  ITEM9
                      |  ITEM14
                      |  ITEM1
                      |S ITEM12  ITEM2
   -2                 +
                      |
                      |  ITEM17  ITEM4
                      |
                      |
                      |  ITEM10
   -3                 +
                      |
                      |
                      |T ITEM13
                      |
                      |
   -4                 +
               <less>|<frequ>
```

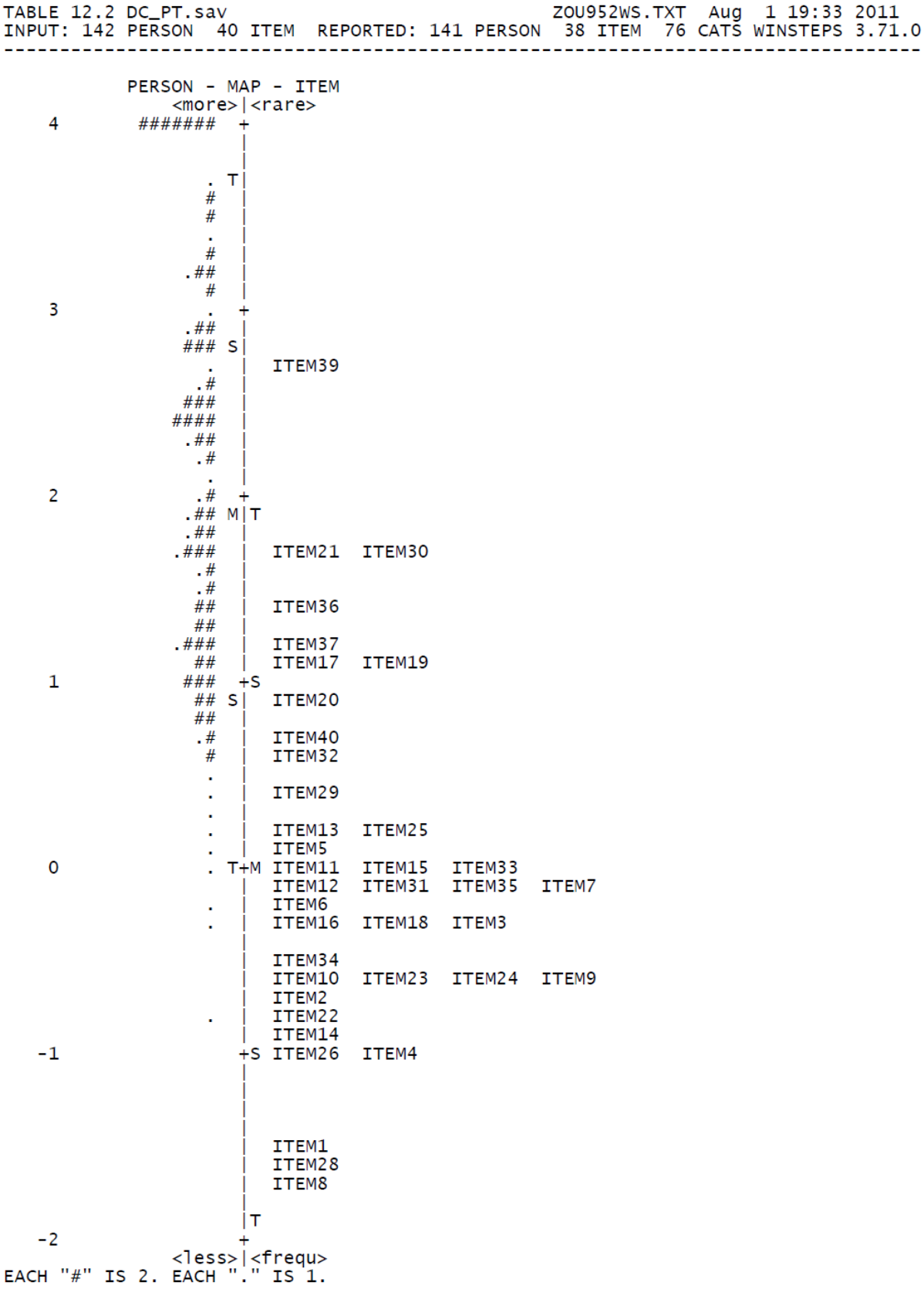
Teste Numérico, versão moçambicana: Mapa de Itens e de Sujeitos

```
TABLE 12.2 NC_MZ.sav                                ZOU088WS.TXT  Aug  1 20:51 2011
INPUT: 418 PERSON  18 ITEM  REPORTED: 418 PERSON  18 ITEM  36 CATS WINSTEPS 3.71.0
-------------------------------------------------------------------------------

           PERSON - MAP - ITEM
               <more>|<rare>
    3    .#########  +
                     |
                     |
                     |
                     |
                   .   |
                   #   |
                       |T
                  .#   |
    2              .   +
                   . T|
                   .   |  ITEM25
                   .   |
                  .#   |
                 ###   |  ITEM28
                 .##   |  ITEM31
                .###   |S ITEM30
    1           .##   +  ITEM29
                .##    |
                .## S  |  ITEM32
                .##    |  ITEM11
                .##    |  ITEM13
                ###    |
              .####    |
              .####    |  ITEM24
    0        .#####   +M ITEM23
              #####    |  ITEM27
                .##    |
             .######   |  ITEM16
              ##### M  |
              .####    |
             ######    |  ITEM18  ITEM21
              .####    |
   -1         .####   +
                .###   |S
             .######   |  ITEM17
              #####    |
                .##    |  ITEM22
                .##    |
                .### S |  ITEM14
                .##    |
   -2        .######  +  ITEM2
                .##    |
                 .#    |T
                 .#    |
                  .    |
                  .    |
                 .#    |
                  # T  |
   -3              .   +
                  .#   |
                   .   |
                       |
                       |
                       |
                       |
   -4              .#  +
                <less>|<frequ>
 EACH "#" IS 3. EACH "." IS 1 TO 2
```

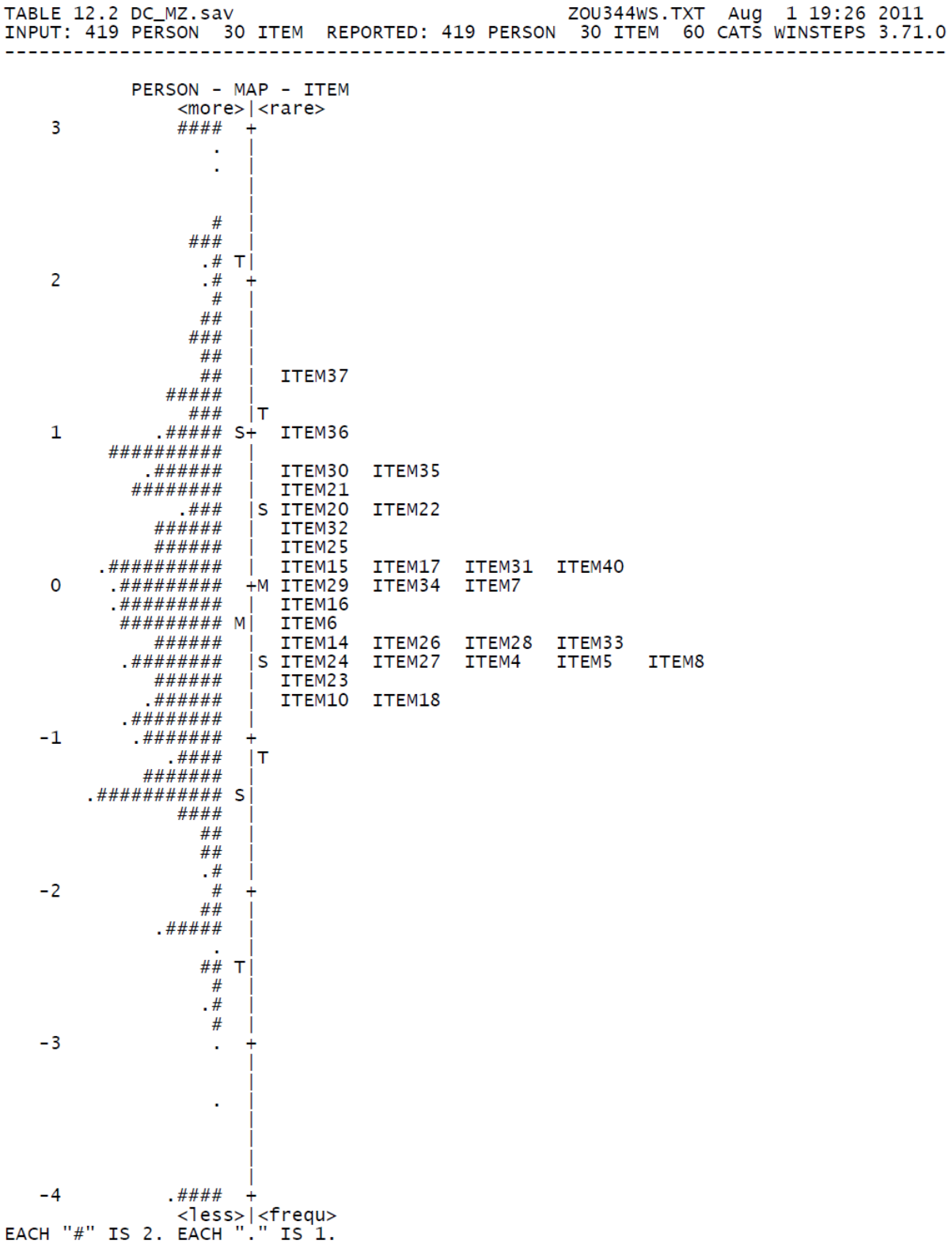
Teste Numérico, versão inglesa: Mapa de Itens e de Sujeitos

```
TABLE 12.2 NC_UK.sav                              ZOU312WS.TXT  Oct  6 18:21 2011
INPUT: 32 PERSON  40 ITEM  REPORTED: 32 PERSON  40 ITEM  80 CATS WINSTEPS 3.71.0
-------------------------------------------------------------------------------

     PERSON - MAP - ITEM
         <more>|<rare>
    6        X  +
                |
                |
                |
                |
          X     |
                |
    5           +
              T |
                |
                |  ITEM36
                |
                |  ITEM37
    4           +
          X     |
                |T
                |  ITEM34
                |
                |
    3      XXX S+  ITEM35
                |
                |
           X    |
          XX    |  ITEM33
          XX    |
    2           +
           X    |S
                |  ITEM26  ITEM31
          XX    |  ITEM11  ITEM22  ITEM28
         XXX    |
                |
    1       XX M+  ITEM15  ITEM20
                |  ITEM23  ITEM29
           X    |  ITEM24
                |  ITEM18  ITEM7
          XX    |
                |  ITEM30
    0     XXXX  +M
                |  ITEM14
           X    |
                |  ITEM32
                |  ITEM2   ITEM21  ITEM3
           X    |  ITEM39
   -1          S+  ITEM5
                |  ITEM13  ITEM19  ITEM25  ITEM27
                |  ITEM17
                |  ITEM16
                |  ITEM38  ITEM4   ITEM8
                |S ITEM40
   -2           +  ITEM12
                |
          XX    |  ITEM1   ITEM10
                |
                |  ITEM6
           X    |
   -3          T+  ITEM9
                |
                |
                |
                |T
           X    |
   -4           +
         <less>|<frequ>
```

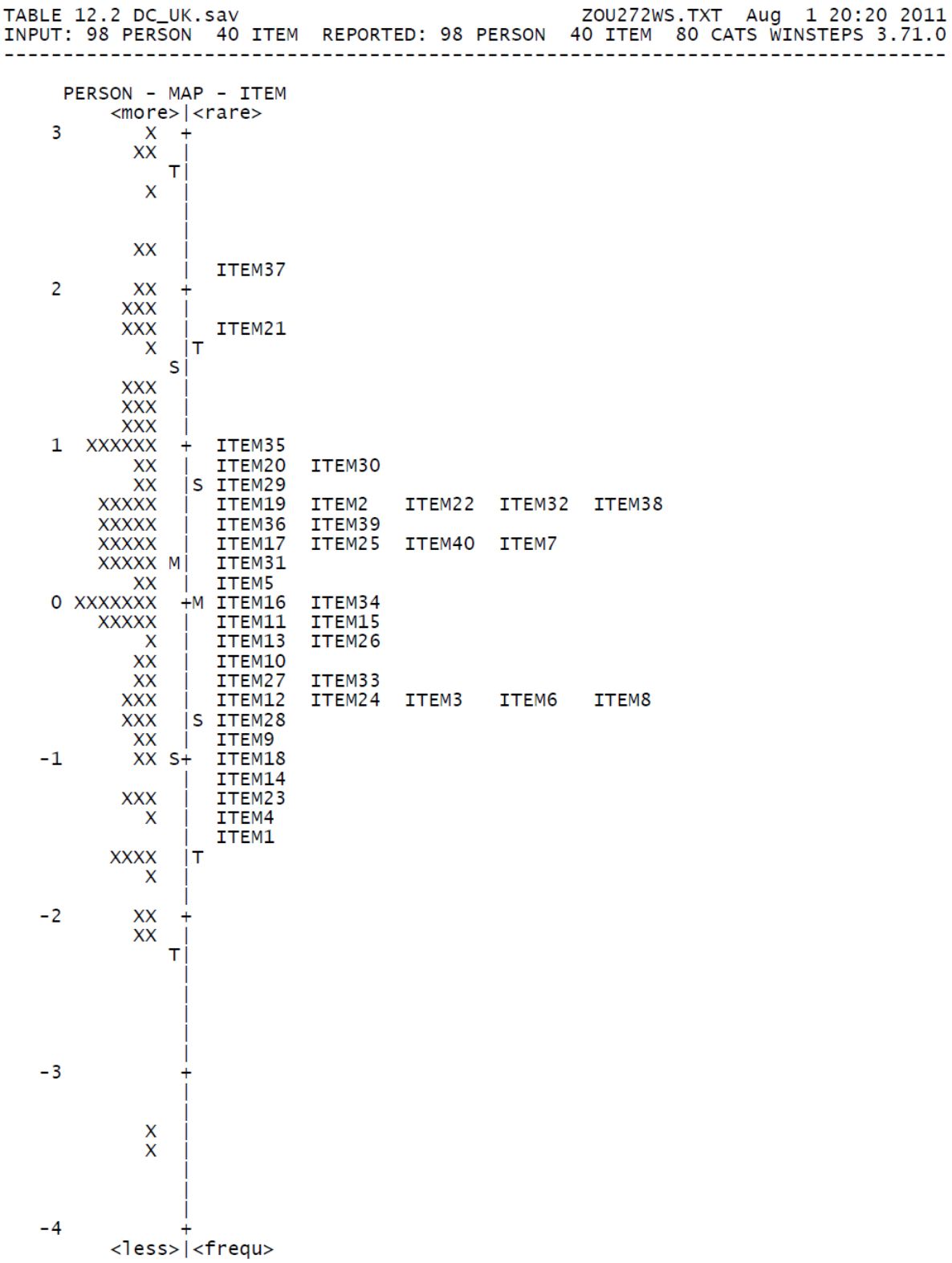
Teste Diagramático, versão portuguesa: Mapa de Itens e de Sujeitos



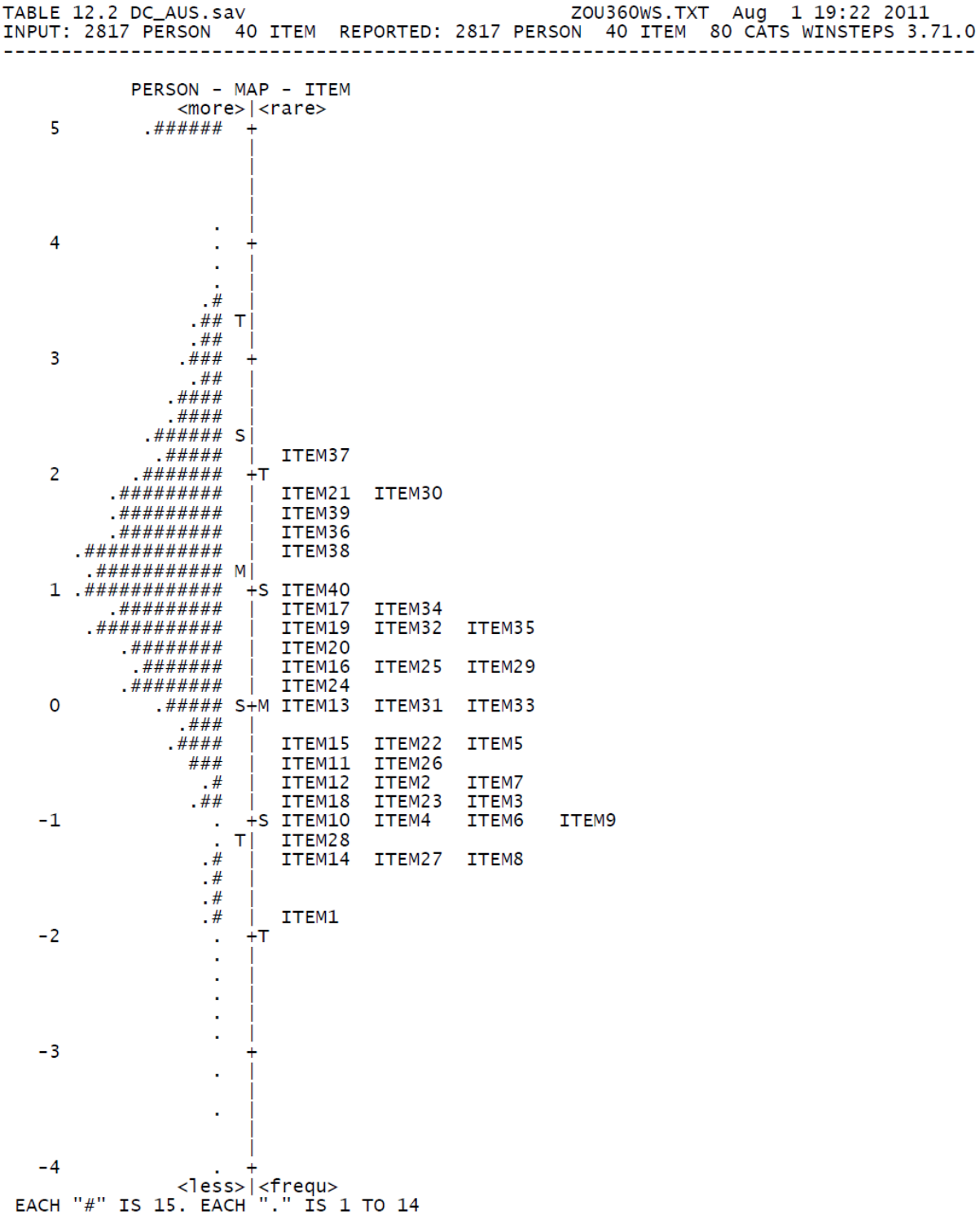
Teste Diagramático, versão moçambicana: Mapa de Itens e de Sujeitos



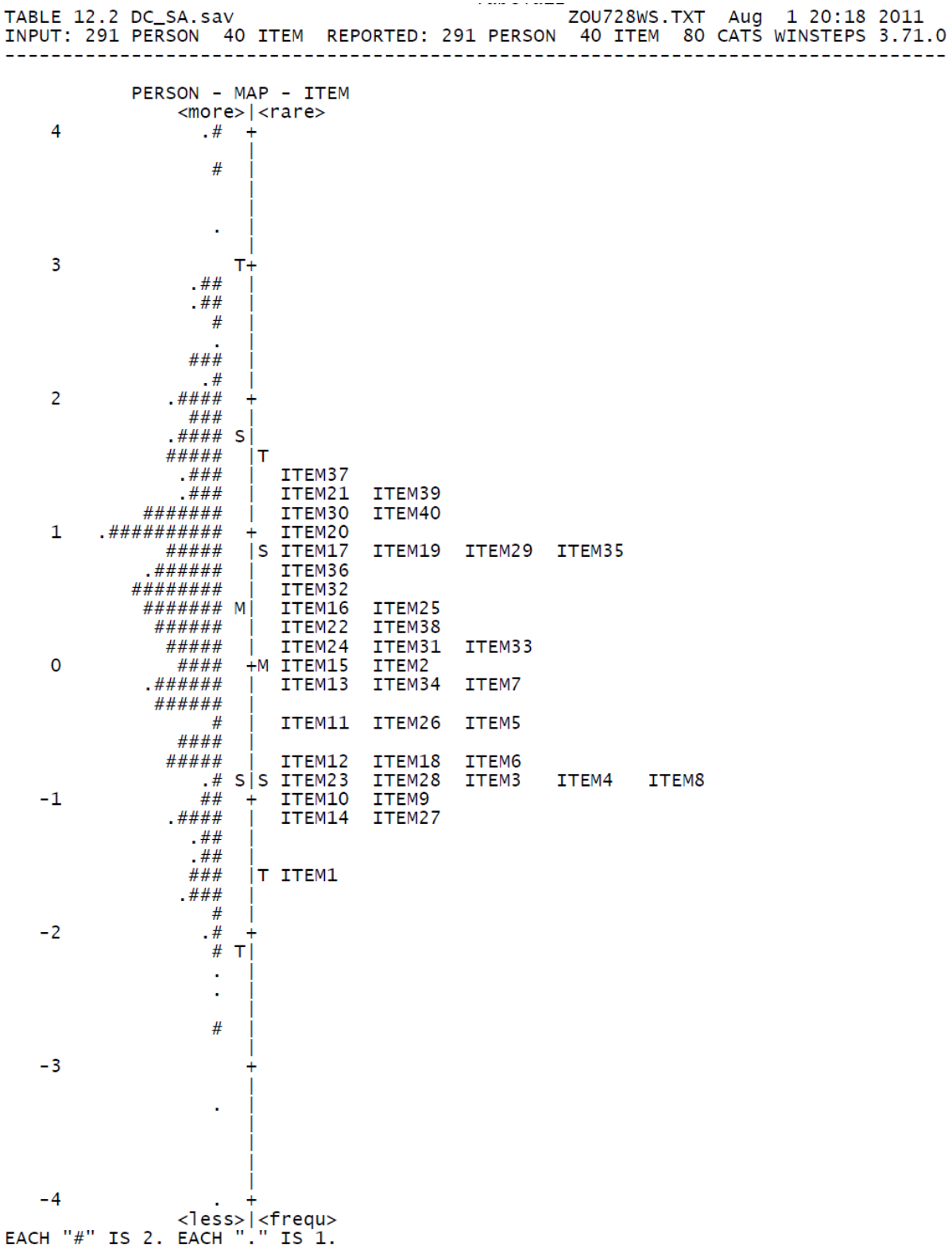
Teste Diagramático, versão inglesa: Mapa de Itens e de Sujeitos



Teste Diagramático, versão australiana: Mapa de Itens e de Sujeitos



Teste Diagramático, versão sul-africana: Mapa de Itens e de Sujeitos



Anexo B

Tabelas de Análise do Funcionamento Diferencial dos Itens (*DIF*)

Teste Verbal: Comparação entre o grupo de referência (Portugal) e o grupo focal (Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF)[1]

Item	Estímulo	Portugal		Moçambique		Contraste *DIF*	*t*	Prob
		Medida *DIF*	se *DIF*	Medida *DIF*	se *DIF*			
1	A	-0.7	.31	-1.01	.12	.32	.96	.3375
2	A	-1.46	.41	-1.80	.15	.34	.77	.4419
4	A	-1.28	.36	-.37	.11	-.91	-2.42	.0165
5	B	.25	.23	.42	.10	-.17	-.69	.4937
6	B	-.61	.30	-.68	.11	.08	.24	.8101
8	B	-.09	.24	.68	.10	-.77	-2.92	.0038
10	C	.40	.21	1.21	.11	-.81	-3.35	.0009
11	C	-.52	.28	.16	.10	-.68	-2.28	.0235
12	C	.25	.23	.31	.10	-.06	-.25	.8012
13	D	-.80	.32	-.66	.11	-.13	-.39	.6943
14	D	-1.29	.38	-.71	.11	-.59	-1.49	.1369
15	D	.53	.21	.93	.11	-.39	-1.68	.0946
17	E	-.52	.28	.14	.17	-.66	-2.00	.0461
20	E	.20	.23	.25	.18	-.06	-.20	.8424
25	F	-.57	.30	-.88	.12	.31	.98	.3302
26	F	-.86	.33	-.42	.11	-.44	-1.27	.2065
33	G	.89	.20	.09	.11	.80	3.48	.0006
34	G	.37	.22	-.47	.12	.84	3.39	.0008
36	G	-.13	.26	-.65	.12	.52	1.85	.0648
37	H	-1.18	.39	-1.54	.16	.36	.87	.3835
38	H	.45	.23	.65	11	-.21	-.83	.4079
40	H	-2.27	.51	-.72	.13	-1.55	-2.98	.0033
41	I	-.08	.26	1.16	.13	-1.24	-4.26	.0000
43	I	.72	.23	.35	.13	.37	1.4	.1635
44	I	1.07	.22	.59	.13	.48	1.87	.0631
45	J	3.64	.28	2.32	.19	1.32	3.91	.0001
46	J	3.52	.25	1.12	.15	2.40	8.26	.0000
54	K	-.98	.56	-1.62	.30	.64	1.00	.3185
55	K	-.01	.41	-.05	.22	.04	.09	.9264
56	K	1.25	.30	.12	.22	1.13	2.99	.0034

[1] Portugal (n=139), Moçambique (n=422)

Teste Numérico: Comparação entre o grupo de referência (Portugal) e o grupo focal (Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF)[2]

Item	Estímulo	Portugal		Moçambique			*t*	Prob
		Medida *DIF*	se *DIF*	Medida *DIF*	se *DIF*	Contraste *DIF*		
2	b	-1.80	.49	-2.03	.14	.23	.45	.6504
11	c	1.13	.26	.63	.15	.50	1.67	.0966
13	e	-3.26	1.01	.50	.16	-3.76	-3.67	.0004
14	f	-1.27	.42	-1.77	.13	.50	1.12	.2639
16	g	-1.23	.43	-.45	.13	-.78	-1.75	.0828
17	d	-2.28	.61	-1.28	.14	-1.00	-1.59	.1135
18	f	-.55	.34	-.75	.13	.21	.57	.5704
21	b	-.10	.32	-.81	.13	.71	2.07	.0398
22	g	-1.05	.50	-1.54	.16	.49	.92	.3580
23	c	-.13	.35	-.05	.16	-.08	-.21	.8309
24	d	.15	.34	.15	.18	.00	.00	1.000
25	f	1.83	.28	1.73	.23	-.09	.26	.7925
27	d	1.07	.33	-.16	.21	1.23	3.17	.0019
28	h	2.28	.37	1.34	.37	.95	1.82	.0716
29	g	1.06	.43	.94	.36	.12	.21	.8324
30	e	.36	.51	1.07	.36	-.71	-1.15	.2553
31	d	1.89	.46	1.27	.52	.62	.90	.3751
32	b	-.02	.60	.69	.36	-.70	-1.01	.3159

[2] Portugal (n=139), Moçambique (n=418)

Teste Diagramático: Comparação entre o grupo de referência (Portugal) e o grupo focal (Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF)[3]

Item	Portugal		Moçambique		Contraste *DIF*	*t*	Prob
	Medida *DIF*	se *DIF*	Medida *DIF*	se *DIF*			
4	-1.10	.36	-.77	.12	-.33	-.87	.3852
5	.08	.25	-.78	.12	.86	3.11	.0021
6	-.28	.27	-.44	.12	.16	.55	.5843
7	-.15	.26	-.26	.12	.11	.38	.7067
8	-1.71	.46	-.72	.12	-.99	-2.07	.0400
10	-.68	.31	-.97	.13	.29	.87	.3878
14	-.99	.34	-.61	.13	-.38	-1.05	.2939
15	-.07	.26	-.12	.13	.04	.15	.8818
16	-.40	.29	-.36	.13	-.04	-.13	.8966
17	1.06	.21	-.06	.14	1.13	4.41	.0000
18	-.36	.28	-1.00	.14	.64	2.03	.0431
20	.84	.21	.33	.14	.52	2.01	.0455
21	1.59	.26	.52	.19	1.07	3.35	.0010
22	-.84	.34	.36	.17	-1.20	-3.15	.0019
23	-.59	.32	-.83	.16	.24	.66	.5086
24	-.62	.35	-.69	.18	.08	.20	.8427
25	.18	.27	.12	.18	.05	.17	.8640
26	-1.04	.43	-.60	.20	-.44	-.92	.3586
27	-4.21	1.80	-.67	.22	-3.54	-1.95	.0540
28	-1.58	.53	-.49	.23	-1.08	-1.89	.0602
29	.40	.27	-.09	.24	.49	-.62	.5331
30	1.69	.30	.61	.34	1.07	2.38	.0191
31	-.13	.36	-.06	.28	-.08	-.17	.8682
32	.63	.32	.26	.33	.38	.82	.4169
33	.00	.39	-.53	.31	.53	1.06	.2925
34	-.44	.55	-.04	.39	-.40	-.58	.5607
35	-.03	.47	.68	.47	-.70	-1.05	.2958
36	1.42	.46	.89	.55	.53	.74	.4652
37	1.27	.53	1.31	.57	-.04	-.05	.9598
40	.79	.60	.03	.45	.75	1.01	.3203

[3] Portugal (n=141), Moçambique (n=419)

Teste Verbal: Comparação entre o grupo de referência (Portugal) e o grupo focal (Reino Unido) para a detecção de Funcionamento Diferencial dos Itens (DIF)[4]

Item	Estímulo	Portugal		Reino Unido		Contraste	*t*	Prob
		Medida *DIF*	se *DIF*	Medida *DIF*	se *DIF*	*DIF*		
5	B	-.01	.23	.06	.49	-.06	-.12	.9049
6	B	-.87	.30	.06	.49	-.93	-1.62	.1111
8	B	-.36	.25	-.47	.54	.12	.19	.8478
9	C	-1.17	.33	-1.67	.77	.51	.60	.5491
11	C	-.78	.29	.29	.47	-1.07	-1.93	.0574
12	C	-.02	.23	.29	.47	-.31	-.59	.5570
17	D	-.78	.29	.54	.48	-1.33	-2.37	.0211
18	D	.60	.20	-.55	.60	1.15	1.83	.0741
19	D	.03	.22	1.44	.46	-1.41	-2.77	.0077
22	L	.85	.19	.06	.49	.79	1.51	.1381
23	L	1.02	.18	.50	.46	.52	1.04	.3025
25	F	-.83	.30	.21	.49	-1.04	-1.81	.0746
26	F	-1.11	.33	.21	.49	-1.33	-2.23	.0288
27	F	.92	.19	.21	.49	.70	1.34	.1880
33	G	.60	.20	-1.12	.66	1.73	2.50	.0169
34	G	.10	.23	-.19	.51	.29	.52	.6054
35	G	3.21	.24	2.05	.45	1.15	2.27	.0267
37	H	-1.42	.39	-.56	.59	-.86	-1.20	.2345
38	H	.18	.22	.77	.47	-.59	-1.13	.2650
41	I	-.34	.28	.31	.50	-.65	-1.14	.2583
42	I	.35	.23	.31	.50	.04	.08	.9371
43	I	.45	.23	-.95	.66	1.40	2.00	.0515
49	M	-.18	.34	-1.52	.77	1.34	1.59	.1179
50	M	-.41	.36	-.32	.54	-.09	-.14	.8874
51	M	-1.22	.52	-.63	.59	-.58	-.74	.4616
52	M	.42	.31	.45	.47	-.02	-.04	.9680

[4] Portugal (n=139), Reino Unido (n=29)

Teste Numérico: Comparação entre o grupo de referência (Portugal) e o grupo focal (Reino Unido) para a detecção de Funcionamento Diferencial dos Itens (DIF)[5]

Item	Estímulo	Portugal		Reino Unido		Contraste *DIF*	*t*	Prob
		Medida *DIF*	se *DIF*	Medida *DIF*	se *DIF*			
1	a	-1.67	.43	-2.32	.80	.66	.72	.4733
2	b	-1.85	.46	-.64	.52	-1.21	-1.76	.0818
3	c	-.86	.32	-.65	.53	-.21	-.34	.7354
4	d	-2.29	.59	-1.51	.61	-.78	-.92	.3587
5	e	.14	.25	-.86	.53	.99	1.70	.0949
6	f	.44	.22	-1.98	.55	2.42	4.07	.0001
7	g	.86	.23	.51	.45	.35	.68	.4961
8	h	-.05	.26	-1.51	.59	1.46	2.28	.0264
9	a	-1.34	.37	-2.81	.84	1.47	1.59	.1169
10	b	-2.84	.71	-2.20	.75	-.64	-.62	.5386
11	c	.75	.23	1.45	.745	-.69	-1.38	.1724
12	d	-1.87	.47	-1.89	.75	.03	.03	.9745
13	e	-3.31	.81	-1.20	.62	-2.11	-2.06	.0419
14	f	-1.42	.40	-.13	.47	-1.29	-2.08	.0402
15	g	-.48	.35	.92	.46	-1.41	-2.43	.0175
16	h	-1.37	.41	-1.41	..65	.003	.04	.9656
17	d	-2.33	.58	-1.31	.64	-1.02	-1.18	.2411
18	f	-.74	.32	.44	.46	-1.17	-2.10	.0394
19	e	-.11	.28	-1.20	.60	1.09	1.65	.1045
20	h	.15	.25	.89	.45	-.73	-1.41	.1635
21	b	-.33	.30	-.74	.58	.41	.63	.5282
22	g	-1.14	.43	1.42	.46	-2.56	-4.07	.0001
23	c	-.38	.32	.79	.46	-1.17	-2.09	.0401
24	d	-.09	.32	.64	.46	-.73	-1.31	.1953
25	f	1.32	.24	-1.00	.52	2.32	4.02	.0002
26	a	2.27	.24	1.61	.45	.65	1.27	.2087
27	d	.58	.29	-1.18	.58	1.76	2.72	.0085
28	h	1.45	.32	1.36	.48	.08	.15	.8845
29	g	.62	.41	.76	.51	-.15	-.22	.8239
30	e	-.17	.47	.18	.54	-.34	-.48	.6337
31	d	1.04	.41	1.60	.49	-.57	-.89	.3788
32	b	-.53	.56	-.53	.63	.00	.00	1.000
33	a	2.32	.37	2.13	.53	.18	.29	.7767
34	g	.78	.67	3.30	.72	-2.51	-2.55	.0168
35	c	1.89	.45	2.73	.71	-.85	-1.01	.3224
36	h	1.52	.72	4.18	.78	-2.66	-2.50	.0196
37	c	5.46	.97	3.95	.68	1.51	1.28	.2095
38	b	2.89	.59	-1.02	.56	4.00	4.89	.0000
39	f	-.35	1.11	-.74	.58	.39	.31	.7588
40	a	-1.81	2.08	-1.70	.67	-.11	-.05	.9608

[5] Portugal (n=139), Reino Unido (n=32)

Teste Diagramático: Comparação entre o grupo de referência (Portugal) e o grupo focal (Reino Unido) para a detecção de Funcionamento Diferencial dos Itens (DIF)[6]

Item	Portugal		Reino Unido		Contraste *DIF*	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
1	-1.64	.43	-1.62	.28	-.01	-.02	.9804
2	-.82	.31	.47	.24	-1.29	-3.26	.0013
3	-.40	.27	-.74	.24	.34	.94	.3473
4	-1.13	.36	-1.47	.27	.34	.75	.4529
5	.04	.25	-.01	.24	.05	.14	.8902
6	-.31	.26	-.77	.25	.46	1.28	.2011
7	-.19	.26	.20	.23	-.39	-1.13	.2583
8	-1.75	.44	-.80	.25	-.95	-1.87	.0624
9	-.70	.31	-1.02	.26	.31	.78	.4352
10	-.72	.30	-.49	.24	-.23	-.60	.5509
11	-.07	.25	-.29	.24	.22	.66	.5105
12	-.23	.26	-.78	.25	.55	1.53	.1279
13	.08	.25	-.35	.24	.43	1.24	.2168
14	-1.03	.34	-1.24	.27	.21	.49	.6222
15	-.11	.26	-.27	.25	.15	.43	.6682
16	-.42	.28	-.13	.25	-.29	-.75	.4527
17	1.03	.21	.25	.24	.78	2.41	.0168
18	-.40	.28	-1.10	.28	.70	1.78	.0774
19	1.01	.21	.56	.26	.45	1.35	.1772
20	.81	.22	.80	.27	.00	.00	.9966
21	1.59	.26	1.68	.35	-.09	-.21	.8303
22	-.86	.33	.54	.27	-1.40	-3.27	.0013
23	-.62	.32	-1.36	.31	.73	1.65	.1003
24	-.64	.33	-.70	.30	.06	.13	.8983
25	.14	.27	.33	.28	-.19	-.49	.6255
26	-1.08	.42	-.30	.30	-.78	-1.53	.1276
27	-3.14	.69	-.55	.29	-2.59	-3.45	.0007
28	-1.62	.51	-.85	.30	-.77	-1.30	.1962
29	.37	.27	.72	.32	-.35	-.84	.4002
30	1.69	.30	.85	.32	.84	1.91	.0583
31	-.17	.36	.23	.32	-.39	-.82	.4136
32	.61	.32	.54	.35	.07	.15	.8786
33	-.05	.40	-.56	.34	.51	.98	.3282
34	-.46	.51	-.10	.35	-.36	-.58	.5645
35	-.10	.46	.90	.41	-1.00	-1.64	.1061
36	1.38	.46	.42	.40	.96	1.57	.1236
37	1.20	.53	2.15	.56	-.95	-1.23	.2261
38	3.27	.98	.56	.45	2.71	2.52	.0306
39	2.72	.61	.45	.45	2.27	2.99	.0053
40	.65	.60	.36	.45	.29	.38	.7034

[6] Portugal (n=141), Reino Unido (n=98)

Teste Diagramático: Comparação entre o grupo de referência (Portugal) e o grupo focal (Austrália) para a detecção de Funcionamento Diferencial dos Itens (DIF)[7]

Item	Portugal		Austrália		Contraste *DIF*	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
1	-1.64	.43	-1.89	.08	.25	.59	.5583
2	-.82	.31	-.66	.05	-.16	-.49	.6231
3	-.40	.27	-.86	.06	.46	1.68	.0954
4	-1.13	.36	-.91	.06	-.22	-.61	.5441
5	.04	.25	-.28	.05	.32	1.27	.2041
6	-.31	.26	-.95	.06	.64	2.40	.0176
7	-.19	.26	-.61	.05	.42	1.60	.1114
8	-1.75	.44	-1.27	.06	-.48	-1.06	.2896
9	-.70	.31	-.94	.06	.24	.76	.4465
10	-.72	.30	-1.06	.06	.34	1.11	.2698
11	-.07	.25	-.47	.05	.40	1.61	.1094
12	-.23	.26	-.66	.05	.43	1.63	.1058
13	.08	.25	.00	.05	.08	.31	.7549
14	-1.03	.34	-1.27	.06	.23	.68	.4990
15	-.11	.26	-.24	.05	.13	.48	.6292
16	-.42	.28	.37	.05	-.79	-2.77	.0064
17	1.03	.21	.92	.05	.11	.52	.6021
18	-.40	.28	-.76	.06	.36	1.25	.2123
19	1.01	.21	.75	.05	.26	1.18	.2387
20	.81	.22	.57	.05	.24	1.09	.2769
21	1.59	.26	1.83	.05	-.24	-.93	.3535
22	-.86	.33	-.31	.05	-.55	-1.64	.1031
23	-.62	.32	-.86	.06	.24	.73	.4654
24	-.64	.33	.23	.05	-.87	-2.60	.0104
25	.14	.27	.33	.05	-.19	-.70	.4855
26	-1.08	.42	-.47	.06	-.61	-1.45	.1504
27	-3.14	.69	-1.26	.07	-1.88	-2.71	.0079
28	-1.62	.51	-1.13	.07	-.49	-.95	.3442
29	.37	.27	.41	.05	-.05	-.17	.8632
30	1.69	.30	1.76	.06	-.07	-.23	.8187
31	-.17	.36	-.04	.06	-.12	-.34	.7332
32	.61	.32	.64	.06	-.04	-.11	.9110
33	-.05	.40	.00	.06	-.05	-.12	.9066
34	-.46	.51	.88	.07	-1.34	-2.60	.0130
35	-.10	.46	.72	.07	-.82	-1.78	.0811
36	1.38	.46	1.42	.07	-.03	-.07	.9437
37	1.20	.53	2.17	.08	-.97	-1.80	.0848
38	3.27	.98	1.35	.08	1.92	1.96	.1077
39	2.72	.61	1.63	.09	1.09	1.77	.0941
40	.65	.60	.36	.45	.29	.38	.7034

[7] Portugal (n=141), Austrália (n=2817)

Teste Diagramático: Comparação entre o grupo de referência (Portugal) e o grupo focal (África do Sul) para a detecção de Funcionamento Diferencial dos Itens (DIF)[8]

Item	Portugal		África do Sul		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
1	-1.64	.43	-1.64	.17	.01	.01	.9895
2	-.82	.31	-.05	.14	-.77	-2.23	.0265
3	-.40	.27	-.91	.15	.51	1.66	.0974
4	-1.13	.36	-.94	.15	-.19	-.49	.6275
5	.04	.25	-.50	.15	.54	1.88	.0608
6	-.31	.26	-.75	.15	.45	1.49	.1377
7	-.19	.26	-.29	.14	.10	.33	.7424
8	-1.75	.44	-.92	.15	-.83	-1.76	.0804
9	-.70	.31	-1.09	.16	.39	1.12	.2618
10	-.72	.30	-1.08	.16	.36	1.07	.2869
11	-.07	.25	-.53	.14	.46	1.63	.1050
12	-.23	.26	-.78	.15	.55	1.82	.0701
13	.08	.25	-.26	.15	.34	1.17	.2432
14	-1.03	.34	-1.27	.16	.23	.62	.5349
15	-.11	.26	-.12	.14	.01	.04	.9688
16	-.42	.28	.33	.15	-.75	-2.00	.0509
17	1.03	.21	.76	.14	.26	1.03	.3026
18	-.40	.28	-.81	.16	.41	1.28	.2002
19	1.01	.21	.80	.15	.21	.81	.4178
20	.81	.22	.86	.15	-.06	-.22	.8240
21	1.59	.26	1.16	.18	.43	1.37	.1720
22	-.86	.33	.28	.15	-1.14	-3.14	.0020
23	-.62	.32	-.86	.17	.24	.66	.5097
24	-.64	.33	.06	.17	-.70	-1.90	.0587
25	.14	.27	.42	.16	-.28	-.90	.3711
26	-1.08	.42	-.50	.17	-.58	-1.29	.1996
27	-3.14	.69	-1.16	.19	-1.98	-2.76	.0066
28	-1.62	.51	-.95	.19	-.66	-1.23	.2225
29	.37	.27	.84	.18	-.47	-1.45	.1479
30	1.69	.30	1.11	.20	.57	1.60	.1131
31	-.17	.36	.13	.19	-.30	-.74	.4621
32	.61	.32	.54	.20	.07	.18	.8553
33	-.05	.40	.11	.20	-.16	-.37	.7112
34	-.46	.51	-.13	.22	-.33	-.60	.5522
35	-.10	.46	.77	.25	-.87	-1.67	.0991
36	1.38	.46	.72	.27	.66	1.24	.2219
37	1.20	.53	1.37	.29	-.17	-.29	.7756
38	3.27	.98	.29	.28	2.98	2.93	.0189
39	2.72	.61	1.28	.32	1.44	2.09	.0454
40	.65	.60	1.07	.34	-.42	-.60	.5496

[8] Portugal (n=141), África do Sul (n=291)

Comparação entre o grupo de referência (Reino Unido) e o grupo focal (Austrália) para a detecção de Funcionamento Diferencial dos Itens (DIF) no Teste Diagramático[9]

Item	Reino Unido		Austrália		Contraste *DIF*	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
1	-1.62	.28	-1.89	.08	.27	.91	.3643
2	.47	.24	-.66	.05	1.13	4.60	.0000
3	-.74	.24	-.86	.06	.12	.47	.6371
4	-1.47	.27	-.91	.06	-.56	-2.01	.0467
5	-.01	.24	-.28	.05	.27	1.11	.2683
6	-.77	.25	-.95	.06	.18	.70	.4863
7	.20	.23	-.61	.05	.81	3.41	.0008
8	-.80	.25	-1.27	.06	.47	1.87	.0635
9	-1.02	.26	-.94	.06	-.07	-.28	.7775
10	-.49	.24	-1.06	.06	.57	2.31	.0222
11	-.29	.24	-.47	.05	.18	.75	.4556
12	-.78	.25	-.66	.05	-.12	-.46	.6429
13	-.35	.24	.00	.05	-.35	-1.44	.1526
14	-1.24	.27	-1.27	.06	.02	.08	.9396
15	-.27	.25	-.24	.05	-.02	-.10	.9216
16	-.13	.25	.37	.05	-.51	-1.96	.0526
17	.25	.24	.92	.05	-.66	-2.69	.0081
18	-1.10	.28	-.76	.06	-.34	-1.21	.2302
19	.56	.26	.75	.05	-.19	-.74	.4587
20	.80	.27	.57	.05	.24	.87	.3868
21	1.68	.35	1.83	.05	-.15	-.43	.6676
22	.54	.27	-.31	.05	.85	3.08	.0027
23	-1.36	.31	-.86	.06	-.49	-1.58	.1178
24	-.70	.30	.23	.05	-.93	-3.03	.0033
25	.33	.28	.33	.05	.00	.00	1.000
26	-.30	.30	-.47	.06	.17	.58	.5667
27	-.55	.29	-1.26	.07	.71	2.37	.0196
28	-.85	.30	-1.13	.07	.28	.90	.3697
29	.72	.32	.41	.05	.30	.95	.3447
30	.85	.32	1.76	.06	-.91	-2.79	.0067
31	.23	.32	-.04	.06	.27	.83	.4085
32	.54	.35	.64	.06	-.11	-.31	.7574
33	-.56	.34	.00	.06	-.56	-1.63	.1082
34	-.10	.35	.88	.07	-.98	-2.72	.0088
35	.90	.41	.72	.07	.18	.43	.6655
36	.42	.40	1.42	.07	-.99	-2.42	.0196
37	2.15	.56	2.17	.08	-.02	-.03	.9742
38	.56	.45	1.35	.08	-.79	-1.71	.0944
39	.45	.45	1.63	.09	-1.18	-2.56	.0142
40	.36	.45	1.04	.09	-.68	-1.49	.1433

[9] Reino Unido (n=98), Austrália (n=2817)

Teste Diagramático: Comparação entre o grupo de referência (Reino Unido) e o grupo focal (África do Sul) para a detecção de Funcionamento Diferencial dos Itens (DIF)[10]

Item	Reino Unido		África do Sul		Contraste *DIF*	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*			
1	-1.62	.28	-1.64	.17	.02	.06	.9552
2	.47	.24	-.05	.14	.52	1.87	.0627
3	-.74	.24	-.91	.15	.17	.60	.5523
4	-1.47	.27	-.94	.15	-.53	-1.69	.0928
5	-.01	.24	-.50	.15	.49	1.75	.0819
6	-.77	.25	-.75	.15	-.01	-.05	.9606
7	.20	.23	-.29	.14	.49	1.80	.0737
8	-.80	.25	-.92	.15	.13	.44	.6635
9	-1.02	.26	-1.09	.16	.07	.25	.8053
10	-.49	.24	-1.08	.16	.60	2.07	.0392
11	-.29	.24	-.53	.14	.24	.87	.3865
12	-.78	.25	-.78	.15	.00	-.01	.9881
13	-.35	.24	-.26	.15	-.09	-.32	.7487
14	-1.24	.27	-1.27	.16	.02	.07	.9469
15	-.27	.25	-.12	.14	-.14	-.50	.6169
16	-.13	.25	.33	.15	-.46	-1.59	.1141
17	.25	.24	.76	.14	-.51	-1.81	.0712
18	-1.10	.28	-.81	.16	-.29	-.90	.3683
19	.56	.26	.80	.15	-.24	-.82	.4123
20	.80	.27	.86	.15	-.06	-.19	.8471
21	1.68	.35	1.16	.18	.52	1.33	.1847
22	.54	.27	.28	.15	.26	.84	.4004
23	-1.36	.31	-.86	.17	-.49	-1.41	.1596
24	-.70	.30	.06	.17	-.76	-2.21	.0290
25	.33	.28	.42	.16	-.09	-.27	.7868
26	-.30	.30	-.50	.17	.20	.59	.5561
27	-.55	.29	-1.16	.19	.61	1.77	.0787
28	-.85	.30	-.95	.19	.11	.29	.7685
29	.72	.32	.84	.18	-.12	-.33	.7411
30	.85	.32	1.11	.20	-.27	-.71	.4810
31	.23	.32	.13	.19	.10	.26	.7967
32	.54	.35	.54	.20	.00	-.01	.9940
33	-.56	.34	.11	.20	-.67	-1.73	.0873
34	-.10	.35	-.13	.22	.03	.07	.9474
35	.90	.41	.77	.25	.13	.28	.7834
36	.42	.40	.72	.27	-.30	-.62	.5370
37	2.15	.56	1.37	.29	.77	1.22	.2248
38	.56	.45	.29	.28	.27	.51	.6137
39	.45	.45	1.28	.32	-.83	-1.51	.1367
40	.36	.45	1.07	.34	-.71	-1.25	.2154

[10] Reino Unido (n=98), África do Sul (n=291)

Teste Diagramático: Comparação entre o grupo de referência (Reino Unido) e o grupo focal (Moçambique) para a detecção de Funcionamento Diferencial dos Itens (DIF)[11]

Item	Reino Unido		África do Sul		Contraste	*t*	Prob
	Medida *DIF*	s.e. *DIF*	Medida *DIF*	s.e. *DIF*	*DIF*		
4	-1.10	.36	-.77	.12	-.75	-2.47	.0144
5	-1.52	.28	-.78	.12	.80	2.90	.0042
6	.02	.24	-.44	.12	-.35	1.25	.2133
7	-.79	.25	-.26	.12	.48	1.82	.0706
8	.22	.24	-.72	.12	-.10	-.37	.7120
10	-.82	.25	-.97	.13	.47	1.68	.0945
14	-.50	.25	-.61	.13	-.68	-2.27	.0244
15	-1.29	.27	-.12	.13	-.17	-.59	.5549
16	-.29	.25	-.36	.13	.22	.76	.4469
17	-.14	.26	-.06	.14	.33	1.16	.2462
18	.27	.25	-1.00	.14	-.16	-.52	.6035
20	-1.16	.28	.33	.14	.44	1.43	.1551
21	.77	.27	.52	.19	1.12	2.81	.0057
22	1.63	.35	.36	.17	.15	.46	.6428
23	.51	.27	-.83	.16	-.62	1.76	.0799
24	-1.44	.31	-.69	.18	-.09	-.25	.8060
25	-.78	.30	.12	.18	.18	.53	.5950
26	.30	.29	-.60	.20	.21	.57	.5674
27	-.39	.30	-.67	.22	.02	.05	.9569
28	-.65	.29	-.49	.23	-.45	1.18	.2403
29	-.94	.31	-.09	.24	.76	1.88	.0625
30	.66	.32	.61	.34	.19	.40	.6920
31	.80	.33	-.06	.28	.21	.49	.6257
32	.16	.32	.26	.33	.21	.43	.6647
33	.47	.35	-.53	.31	-.12	-.26	.7992
34	-.65	.34	-.04	.39	-.13	-.25	.8009
35	-.17	.36	.68	.47	.16	.26	.7974
36	.84	.41	.89	.55	-.52	-.76	.4511
37	.37	.41	1.31	.57	.82	1.02	.3135
40	2.13	.57	.03	.45	.24	.38	.7070

[11] Reino Unido (n=98), Moçambique (n=419)